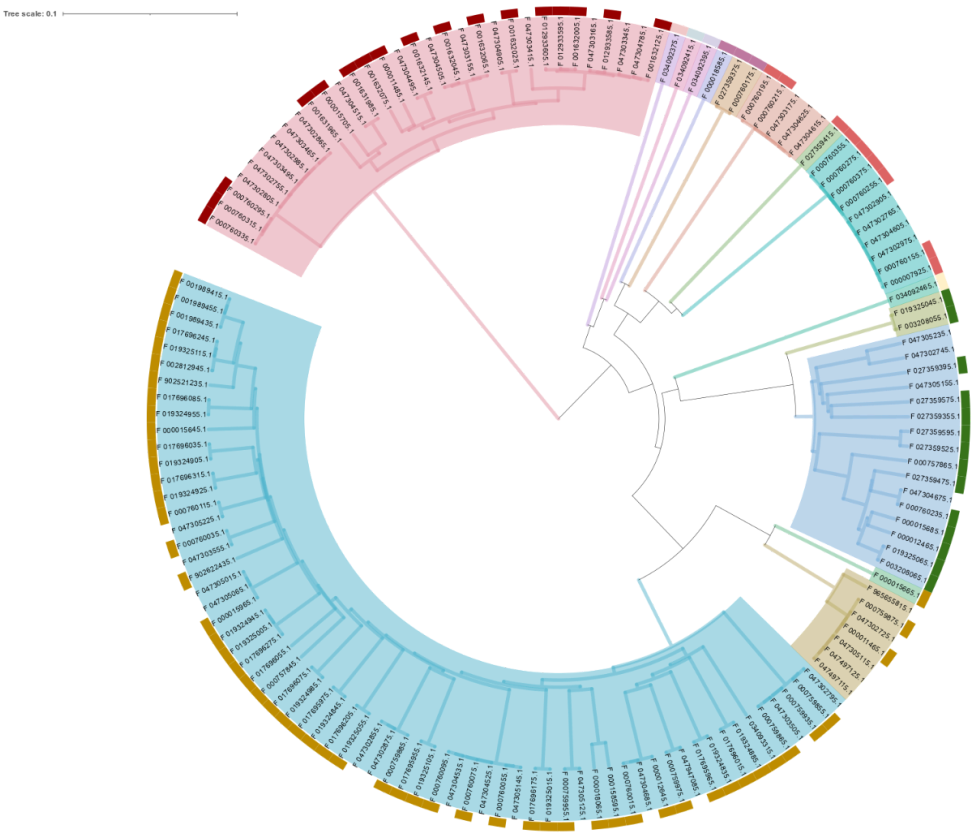
Étude comparative du génome et de l'évolution de Prochlorococcus à partir de données RefSeq et Zhang et al. (2021)

Phylogénomique

Rapport de projet
Effectué dans le cadre du parcours
Master 2 Bioinformatique et Biologie des Systèmes
Année universitaire 2024-2025



Arbre phylogénétique des génomes de Prochlorococcus visualisé à l'aide d'iTOL
(Figure 30 des données supplémentaires)

Université de Toulouse
Projet à rendre le 24/10/2025
Étudiant : RODRIGUÈS Camille-Astrid
Enseignant UE : QUENTIN Yves

Résumé

Le collectif *Prochlorococcus*, principal représentant des cyanobactéries marines photosynthétiques, constitue un modèle de choix pour l'étude de l'adaptation et de la diversification microbienne à l'échelle génomique. Ce projet visait à caractériser la diversité génétique et phylogénomique de *Prochlorococcus* à partir de jeux de données issus des bases publiques RefSeq et des travaux de Zhang et al. (2021). Après filtrage et contrôle qualité des génomes (complétude, contamination), un sous-échantillon représentatif a été retenu afin de réduire la redondance et d'optimiser les temps de calcul.

L'annotation génomique a été réalisée avec **Prokka**, suivie d'une annotation fonctionnelle via **eggNOG-mapper**. L'inférence des groupes d'orthologues avec **OrthoFinder** a permis de distinguer le génome cœur, le génome accessoire et les gènes spécifiques, et de calculer les événements de duplication. Les analyses de pan-génome, effectuées sous **R** à l'aide du script **panplots.R**, ont montré un **pan-génome ouvert** et un **génome cœur stable**, témoignant d'une forte plasticité génétique et d'un noyau fonctionnel conservé.

Les alignements multiples de gènes conservés, évalués par **T-Coffee/TCS**, ont servi à la reconstruction d'arbres d'espèces à l'aide d'**IQ-TREE,** selon différents modèles de substitution (nucléotidique, codonique et protéique). Ces résultats ont été comparés à des super-arbres construits par les approches **MRP**, **ASTRAL** et **consensus majoritaire**, toutes convergeant vers une topologie cohérente distinguant clairement les écotypes de **haute lumière (HL)** et de **basse lumière (LL).**

L'ensemble des analyses met en évidence une **structuration phylogénétique robuste**, une **stabilité du génome cœur** et une **variabilité fonctionnelle importante** entre souches. Ce travail illustre la mise en œuvre d'une approche intégrée et reproductible de phylogénomique comparative sur la plateforme **GenoToul**, combinant rigueur bioinformatique et exploration évolutive, et contribue à mieux comprendre les processus de différenciation et d'adaptation écologique au sein du collectif *Prochlorococcus*.

Points clés

- Pipeline complet de phylogénomique : mise en œuvre d'un flux d'analyse reproductible, depuis la constitution du jeu de données jusqu'à la construction des arbres d'espèces.
- Pan-génome ouvert : courbes d'accumulation (panplots) révélant une diversité génétique croissante et un génome cœur stable.
- Approches phylogénétiques multiples : arbres obtenus par super-alignement (IQ-TREE) et super-arbres (MRP, ASTRAL, consensus) cohérents entre eux.
- Intégration et reproductibilité : analyses effectuées sur la plateforme GenoToul avec scripts Bash et R documentés.

Mots-clés

Prochlorococcus, phylogénomique, pan-génome, orthogroupes, génome cœur, génome accessoire, OrthoFinder, IQ-TREE, ASTRAL, MRP, annotation génomique, Prokka, eggNOG-mapper, diversité génétique, écotypes HL/LL, cyanobactéries marines, adaptation écologique, bioinformatique comparative, pipeline reproductible, GenoToul.

Introduction

La phylogénomique est une discipline à l'interface entre la génomique comparative et la phylogénie moléculaire. Elle vise à étudier les relations évolutives entre espèces à partir de l'analyse de leurs génomes complets.

En intégrant des outils de bioinformatique et de modélisation, elle permet d'explorer la structure, la composition et la dynamique des génomes afin de retracer l'histoire évolutive des lignées et d'identifier les mécanismes d'adaptation.

L'importante littérature écologique consacrée au collectif *Prochlorococcus* repose traditionnellement sur l'hypothèse qu'il s'agit d'un **unique genre**, représenté par l'espèce *Prochlorococcus marinus*, elle-même constituée d'un ensemble d'écotypes. Selon cette vision, la **diversité génomique** observée au sein du collectif s'explique par le concept de **pan-génome ouvert**, où le répertoire génétique global est distribué entre les différentes souches.

Cependant, l'arrivée de **nouvelles données génomiques à haute résolution** a conduit les chercheurs à **réévaluer la classification taxonomique** de ce groupe à l'aide des méthodes modernes de taxonomie génomique.

Dans le cadre de ce projet, nous avons étudié la diversité génomique du genre *Prochlorococcus*, un cyanobactérie marine photosynthétique représentant l'un des organismes les plus abondants de la planète. Ces organismes photosynthétiques, particulièrement abondants dans les océans, représentent un modèle de choix pour étudier la diversification adaptative et les phénomènes de réduction génomique. En effet, *Prochlorococcus* présente une forte variabilité du contenu génétique, reflet de son adaptation à différents régimes lumineux et conditions environnementales.

Jeux de données et prétraitements

Les génomes ont été sélectionnés à partir des ensembles publiés par *Zhang et al. (2021), Kettler et al. (2007)* et *Yan et al. (2018).* Après téléchargement depuis la base RefSeq, les séquences ont été filtrées et renommées pour assurer la compatibilité des identifiants. Les analyses ont été effectuées sur la plateforme de calcul **GenoToul** à l'aide de scripts **Bash** et **R**, en utilisant le gestionnaire de tâches **Slurm**.

Logiciels et outils utilisés

Plusieurs logiciels spécialisés ont été mobilisés au cours des différentes étapes du pipeline :

Etape	Logiciel principal	Objectif
Annotation génomique	**Prokka**	Identifier et annoter les gènes codants, ARNr et ARNt à partir des séquences brutes.
Annotation fonctionnelle	**eggNOG-mapper**	Associer une fonction biologique et une catégorie COG aux protéines prédites.
Inférence d'orthologues	**OrthoFinder**	Regrouper les gènes en orthogroupes et identifier les duplications, pertes et orthologues uniques.
Analyse pan-génomique	**R (ggplot2, matrixStats)**, **panplots**	Évaluer la taille du pan-génome et du génome cœur selon la loi de Heaps.
Sélection des familles conservées	**T-Coffee / Transitive Consistency Score (TCS)**	Évaluer la qualité des alignements multiples et filtrer les alignements les plus fiables.
Alignements et inférence phylogénétique	**IQ-TREE, ASTRAL, phytools (R)**	Construire des arbres phylogénétiques à partir des alignements concaténés ou individuels (super-matrice / super-arbre).
Visualisation et analyses statistiques	**R (ape, ggplot2, TreeTools, ggtree)**	Représenter, comparer et interpréter les arbres espèces.

Analyses réalisées

Les analyses ont suivi une démarche progressive et reproductible :

1. **Téléchargement et sélection des génomes** : constitution de deux ensembles (*Prochlorococcus* et *Synechococcus*) à partir des travaux de *Zhang et al.*
2. **Annotation génomique avec Prokka** : standardisation des prédictions de gènes et des séquences protéiques.
3. **Annotation fonctionnelle avec eggNOG-mapper** : classification fonctionnelle selon les catégories COG.
4. **Inférence d'orthogroupes avec OrthoFinder** : identification des gènes homologues, des duplications et des orthologues uniques.
5. **Analyse du pan-génome** : estimation du génome cœur, du génome accessoire et des gènes spécifiques selon la loi de Heaps.
6. **Sélection de familles conservées** : extraction de gènes du génome cœur sans paralogues et évaluation des alignements multiples par T-Coffee/TCS.
7. **Construction d'arbres phylogénétiques** :
    a. par **super-alignement** (IQ-TREE, modèles nucléotidique, codon et protéique),
    b. par **super-arbre** (MRP, consensus, ASTRAL).
8. **Comparaison des topologies d'arbres** : calcul des distances de Robinson-Foulds pour mesurer la similarité entre méthodes et modèles.

L'ensemble de ces analyses vise à mieux comprendre les relations phylogénétiques entre les clades de *Prochlorococcus* et leurs proches parents *Synechococcus*, ainsi qu'à explorer l'impact de la réduction génomique et des événements de duplication ou de perte de gènes dans leur évolution. Au-delà de l'étude biologique, ce travail illustre la mise en œuvre intégrée de méthodes et d'outils bioinformatiques modernes (Bash, R, Slurm, Prokka, OrthoFinder, IQ-TREE, ASTRAL) sur les infrastructures de calcul **GenoToul**, dans une démarche reproductible et rigoureuse de phylogénomique comparative.

Création d'un jeu de données

La première étape de ce travail était de constituer un **jeu de données génomiques complet** destiné à servir de base à l'étude comparative du collectif *Prochlorococcus*. Cette phase de construction avait pour but d'intégrer l'ensemble des génomes disponibles dans les bases publiques, afin d'explorer la diversité du groupe à grande échelle. Toutefois, en raison du nombre élevé de génomes potentiellement mobilisables et du temps de calcul associé aux analyses pan-génomiques et phylogénomiques, un **sous-échantillon représentatif**— issu de l'étude de *Zhang et al. (2021)* — a été sélectionné pour la suite des analyses. Ce jeu de données réduit permet de conserver la diversité génétique et écologique essentielle tout en rendant le traitement bioinformatique plus rapide et plus maîtrisable.

Constitution du jeu de données complet

La création initiale du jeu de données a reposé sur le **téléchargement des génomes complets** des genres *Prochlorococcus* et *Synechococcus* depuis la base de données **RefSeq** du **NCBI**. Ces données publiques présentent l'avantage d'être vérifiées, standardisées et régulièrement mises à jour. Chaque génome téléchargé est accompagné d'un fichier de métadonnées comportant des informations essentielles telles que le numéro d'accession, le nom de la souche, l'espèce, le type d'échantillon et le lien FTP permettant d'accéder directement à la séquence d'ADN complète. Ces métadonnées ont été extraites, nettoyées et formatées de manière à être compatibles avec les traitements automatisés réalisés ultérieurement.

Une première phase de **filtrage des génomes** a ensuite été mise en œuvre afin d'assurer la qualité du jeu de données. Comme l'illustre la Figure 1, une contamination du jeu de données peut avoir lieu lors de la récupération des données. La recherche de génomes de taille aberrante et leur identification a permis de nettoyer le jeu de données des génomes de phage de *Prochlorococcus* présents dans la base.

Les critères retenus pour le filtrage ont porté sur la complétude de l'assemblage (Fig.2), l'absence d'erreurs d'annotation majeures. Les génomes incomplets, redondants ou supprimés de RefSeq ont été exclus. Cette étape garantit la fiabilité des comparaisons et limite les biais susceptibles d'affecter les inférences évolutives.

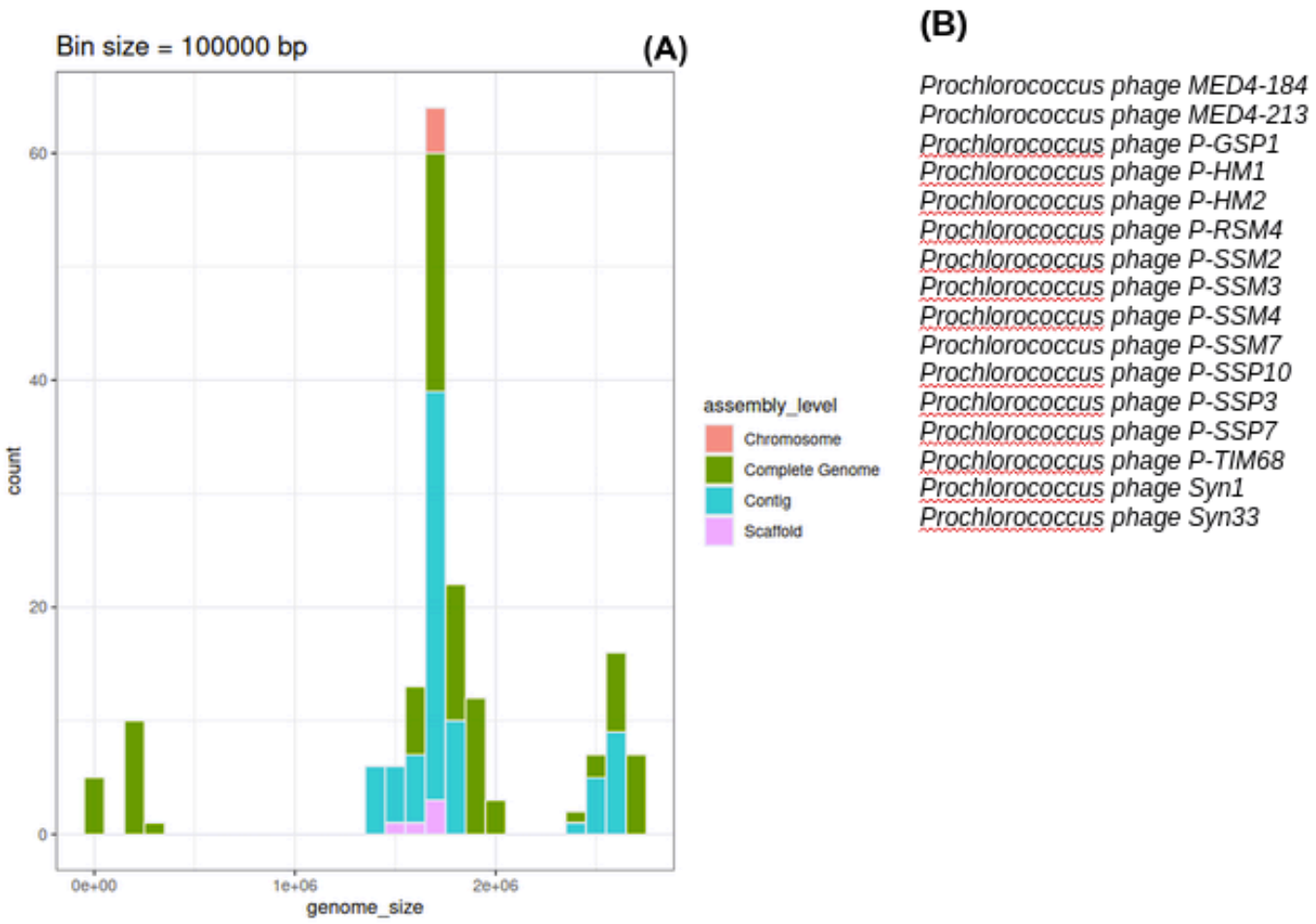


Figure 1. (A) Distribution des tailles de génomes chez les souches de *Prochlorococcus* issues de RefSeq. La figure met en évidence une variabilité de la taille génomique entre les génomes observés. Les petits génomes semblent ne pas appartenir aux *Prochlorococcus.* **(B) Analyse des métadonnées appartenant à ces génomes.** Le résultat de l'analyse montre que les petits génomes identifiés en Fig1.A correspondent à une contamination du jeu de données par des génomes de phage de *Prochlorococcus.*

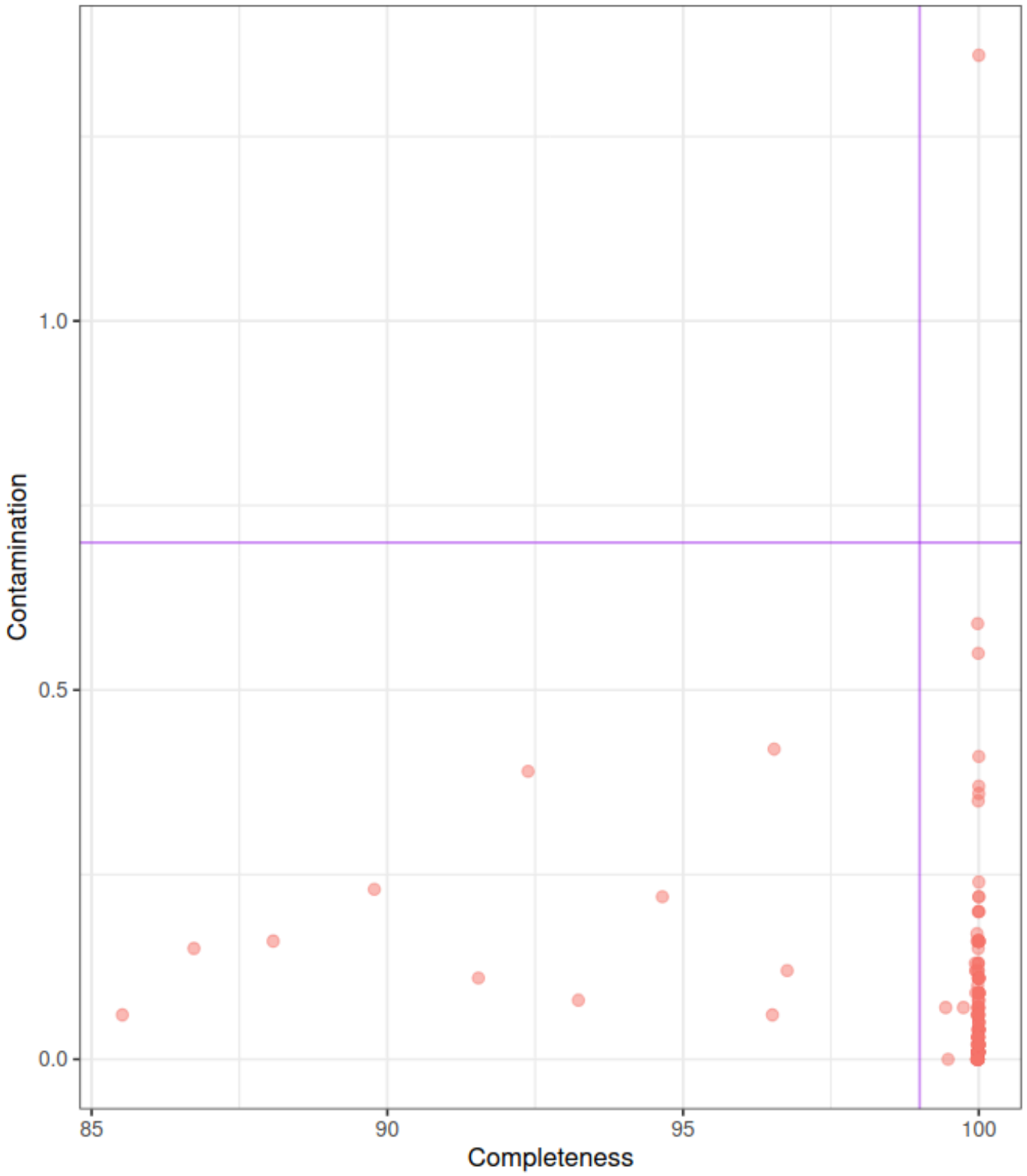


Figure 2. Nuage de points représentant la relation entre la complétude et la contamination des génomes de *Prochlorococcus* évaluées avec CheckM. La figure illustre la forte proportion de génomes présentant une complétude élevée et une contamination minimale. La présence de 12 génomes non conforme à la qualité attendue (contamination < 50% et complétude > 98%) seront retirés de la suite de l'analyse

Une fois les génomes filtrés, un système de **codification interne** a été établi afin de faciliter le suivi des souches tout au long de l'analyse. Chaque génome s'est vu attribuer un code court et unique, composé d'un préfixe indiquant le genre (*Pr* pour *Prochlorococcus*, *Sy* pour *Synechococcus*) et de lettres identifiant la souche. Cette convention de nommage a permis de standardiser les fichiers en entrée des logiciels utilisés, d'automatiser les étapes de traitement et de garantir la reproductibilité des analyses.

Sélection du sous-échantillon de Zhang

Pour les étapes expérimentales suivantes (annotation, inférence d'orthologues, analyses pan-génomiques et phylogénétiques), un sous-ensemble restreint de génomes a été sélectionné à partir du travail de *Zhang et al. (2021)*.

Ce choix répondait à un double objectif : **réduire le temps de calcul** nécessaire sur l'infrastructure de calcul partagée et **faciliter l'interprétation comparative** avec les résultats publiés. L'échantillon de Zhang regroupe quarante génomes de *Prochlorococcus* et vingt de *Synechococcus*, couvrant les principaux écotypes identifiés. Ces écotypes se distinguent notamment par leurs adaptations physiologiques à la lumière, à la température et à la disponibilité en nutriments. Ainsi, le sous-échantillon conserve la **représentativité écologique et phylogénétique** du collectif tout en limitant la redondance et le volume de données à traiter.

Avant utilisation, les fichiers de métadonnées issus de l'étude de *Zhang et al. (2021)* ont été harmonisés avec les jeux RefSeq complets. Les identifiants de génomes ont été standardisés, les doublons supprimés et les liens FTP mis à jour. Ce travail d'uniformisation était indispensable pour garantir la compatibilité entre les différentes sources de données et permettre leur exploitation conjointe dans les étapes d'annotation automatique et de comparaison d'orthologues.

Environnement et organisation des données

L'ensemble des traitements a été réalisé sur la **plateforme de calcul GenoToul** (INRAE, Toulouse), au sein d'un environnement Linux utilisant le **gestionnaire de ressources Slurm** pour le lancement et le suivi des tâches. Les scripts de préparation, d'importation et de filtrage des données ont été développés en **Bash** et en **R**, ce qui a permis de combiner la flexibilité du traitement de texte avec la rigueur des manipulations statistiques.

Finalité du jeu de données

Le jeu de données ainsi obtenu offre une **base solide et homogène** pour les analyses suivantes. Sa construction progressive — depuis le jeu complet issu de RefSeq jusqu'à l'échantillon restreint de Zhang — permet d'allier **exhaustivité et efficacité** : le premier garantit la validité biologique et la couverture du groupe étudié, tandis que le second rend possibles les analyses phylogénomiques détaillées dans des délais raisonnables.

Cette approche itérative illustre l'un des principes fondamentaux de la bioinformatique comparative : adapter la taille du jeu de données aux ressources de calcul disponibles et à la profondeur d'analyse souhaitée, sans compromettre la qualité ni la représentativité scientifique du matériel génétique étudié.

Annotation des génomes

L'annotation génomique constitue une étape essentielle dans l'analyse phylogénomique. Elle vise à identifier, sur les séquences d'ADN, les régions codantes, les gènes fonctionnels et les éléments non codants. Dans ce projet, l'annotation a été réalisée à l'aide du logiciel **Prokka**, un outil largement utilisé pour l'annotation rapide et automatisée des génomes bactériens.

Prokka s'appuie sur un ensemble de programmes spécialisés, tels que **Prodigal** pour la détection des gènes codants, **RNAmmer** et **Aragorn** pour l'identification des ARN ribosomiques et de transfert, et **SignalP** ou **Infernal** pour la prédiction de peptides signaux et d'ARN non codants. L'outil combine ainsi plusieurs approches complémentaires de reconnaissance de motifs et d'alignement de séquences, garantissant une annotation fiable et cohérente.

L'annotation a été effectuée sur l'ensemble des génomes issus du sous-échantillon de Zhang, couvrant les genres *Prochlorococcus* et *Synechococcus*. Pour chaque génome, Prokka a produit un ensemble de fichiers standardisés : les annotations au format GenBank, les fichiers GFF pour les analyses génétiques, ainsi que les séquences d'ADN codant (ffn) et de protéines (faa). Ces résultats ont permis d'obtenir une description complète et homogène du contenu génétique des souches analysées.

L'intérêt principal de Prokka réside dans sa **rapidité**, sa **standardisation** et sa **compatibilité** avec les étapes suivantes de l'analyse, notamment l'inférence d'orthologues. Cependant, ses limites résident dans sa dépendance à des bases de données internes qui peuvent ne pas être exhaustives, et dans sa tendance à attribuer l'étiquette « protéine hypothétique » aux gènes pour lesquels aucune correspondance fiable n'est trouvée. Malgré ces limites, Prokka reste une solution de référence pour générer des annotations reproductibles dans des pipelines à grande échelle.

Annotation fonctionnelle

Une fois les gènes prédits, une **annotation fonctionnelle** a été réalisée afin d'attribuer à chaque protéine un rôle biologique et une classification dans des familles ou catégories métaboliques. Cette étape a été menée avec **eggNOG-mapper**, un outil basé sur les groupes orthologues pré-calculés de la base **eggNOG** (evolutionary genealogy of genes: Non-supervised Orthologous Groups).

EggNOG-mapper utilise une approche fondée sur l'homologie et la phylogénie pour transférer des fonctions à de nouvelles séquences en identifiant leurs orthologues dans les bases de données de référence. Cette méthode est plus précise que les recherches par similarité directe (de type BLAST), car elle limite le risque de transférer des annotations issues de paralogues aux fonctions divergentes.

Dans le cadre du projet, les séquences protéiques issues de Prokka ont été soumises à eggNOG-mapper via le module **Diamond**, permettant une recherche rapide d'orthologues dans les bases de données fonctionnelles. Les résultats ont fourni, pour chaque protéine, une annotation fonctionnelle détaillée : nom du produit, catégorie COG, voie métabolique associée et liens vers les bases de données KEGG et GO.

L'intérêt de cette approche réside dans sa capacité à fournir une vision d'ensemble des fonctions génétiques présentes dans chaque génome, facilitant ainsi les analyses de diversité fonctionnelle et les comparaisons entre espèces. Les principales limites tiennent au fait que certaines fonctions demeurent inconnues ou partiellement annotées, notamment pour les gènes spécifiques à certaines lignées, et que la qualité de l'annotation dépend directement de la couverture et de la mise à jour des bases de données d'orthologues.

Obtention des orthogroupes

L'étape suivante a consisté à identifier les **orthologues** entre les différents génomes (Fig.3-4), c'est-à-dire les gènes issus d'un même gène ancestral et partageant une fonction équivalente. Cette analyse a été réalisée à l'aide du logiciel **OrthoFinder**, un outil de référence pour la génomique comparative.

OrthoFinder procède par alignement de toutes les protéines entre elles à l'aide du moteur **DIAMOND**, puis construit un graphe de similarité pondéré dont les sommets représentent les gènes et les arêtes, leurs similarités. Les gènes sont ensuite regroupés en **orthogroupes** à l'aide de l'algorithme de clustering **MCL (Markov Cluster Algorithm)**. Pour chaque orthogroupe, un arbre de gènes est inféré, permettant à OrthoFinder d'en déduire un **arbre d'espèces enraciné** et d'identifier les événements de duplication et de perte de gènes.

Cette approche présente l'avantage d'être **rapide, robuste et phylogénétiquement informée**, tout en fournissant une base solide pour les analyses de pan-génome et les reconstructions évolutives. Ses limites résident dans la dépendance à la qualité des annotations initiales : des erreurs dans la prédiction des gènes peuvent se propager et influencer la composition des orthogroupes. De plus, les duplications récentes (Fig.5) et les transferts horizontaux peuvent parfois brouiller la distinction entre orthologues et paralogues.

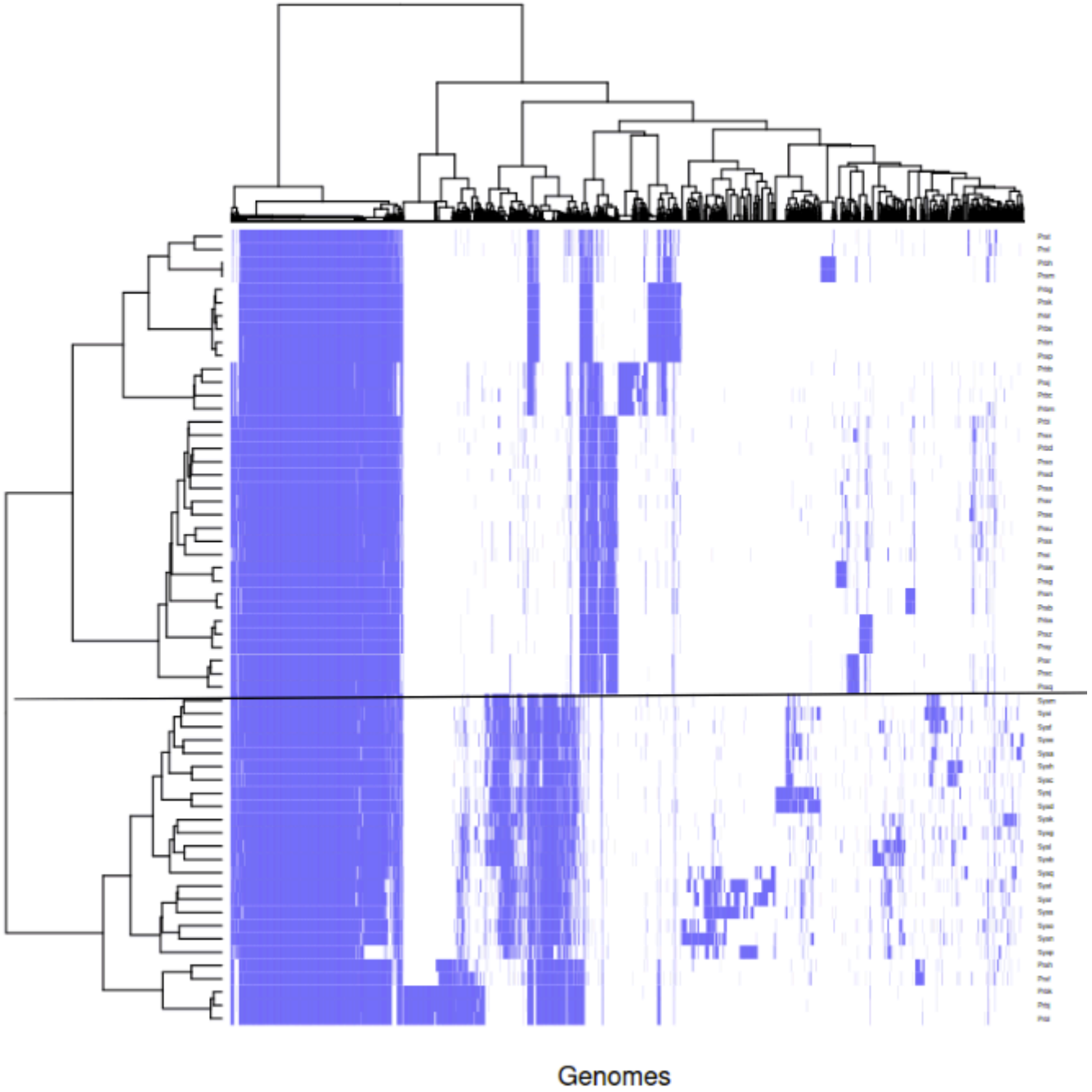


Figure 3. Matrice binaire de présence/absence des orthogroupes identifiés par OrthoFinder pour les souches de *Prochlorococcus* et *Synechococcus*. Chaque colonne représente un orthogroupe et chaque ligne une souche ; la figure met en évidence la répartition du génome cœur, des gènes accessoires et des gènes spécifiques reflétant la diversité génétique du collectif. Les souches au-dessus de la séparation (trait noir horizontal) correspondent aux *Prochlorococcus*, tandis que ceux en-dessous correspondent aux *Synechococcus.*
La vérification de la présence/absence des orthogroupes est effectuée par calcul de la distance euclidienn. La présence du HOG dans tous les génomes indique le génome coeur. 95% de présence sera utilisé comme seuil pour la détermination de l'appartenance au génome coeur, les génomes étant incomplets.
L'étude du génome coeur permet de retracer l'arbre espèce, tandis que celle du génome accessoire permet d'identifier les adaptations aux différents milieux.

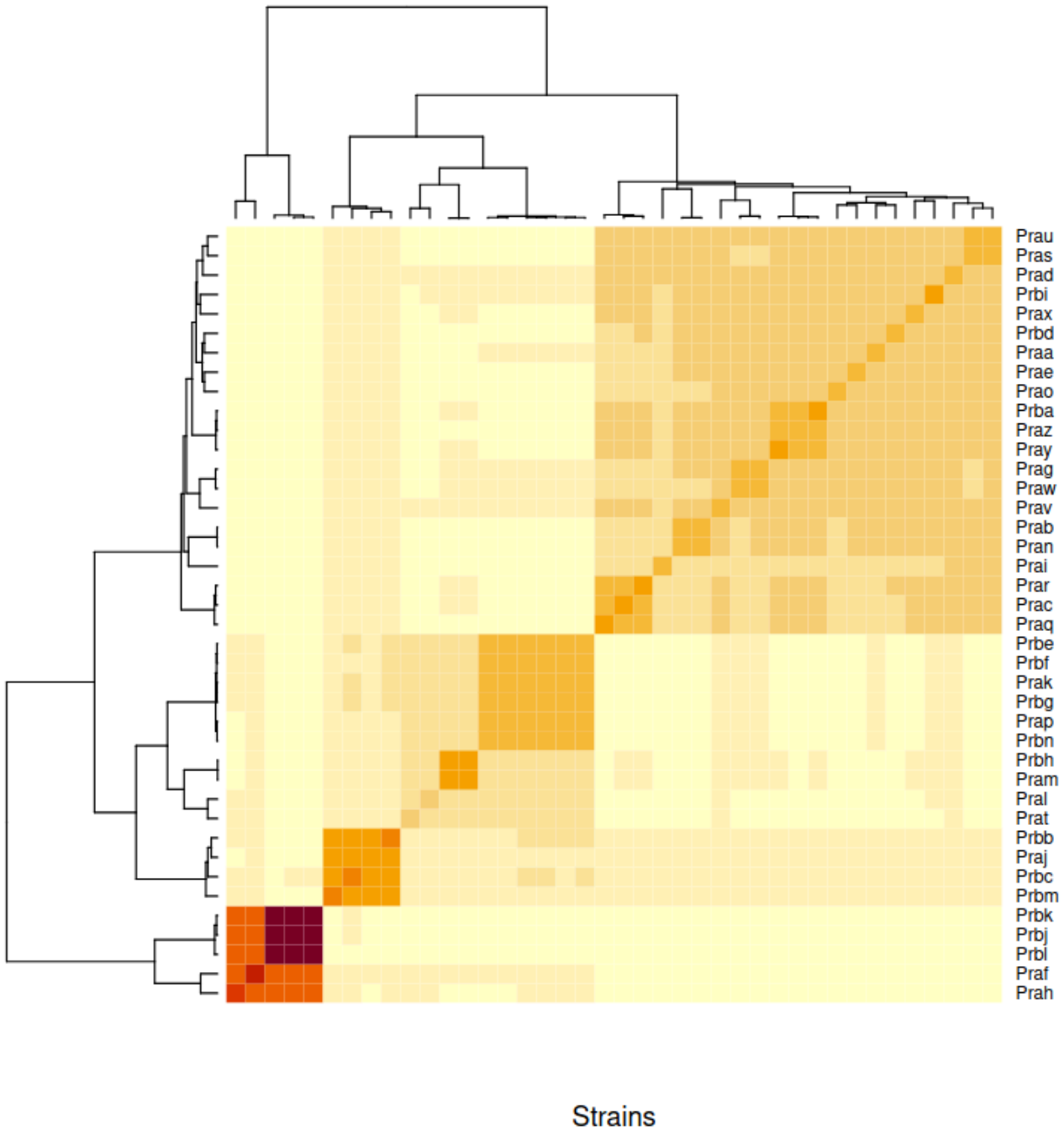


Figure 4. Heatmap représentant les chevauchements d'orthogroupes entre les différentes souches de *Prochlorococcus*, obtenue à partir des résultats d'OrthoFinder. Chaque cellule indique la proportion de gènes partagés entre deux souches, tandis que les dendrogrammes adjacents reflètent la similarité génomique globale. Les zones plus foncées traduisent un partage important de gènes, correspondant aux clades proches phylogénétiquement, alors que les zones plus claires indiquent une divergence génétique accrue. Cette représentation met en évidence l'existence d'un **génome cœur fortement conservé** ainsi qu'un **génome accessoire étendu**, témoignant de la diversité fonctionnelle et de la plasticité adaptative du collectif *Prochlorococcus*.

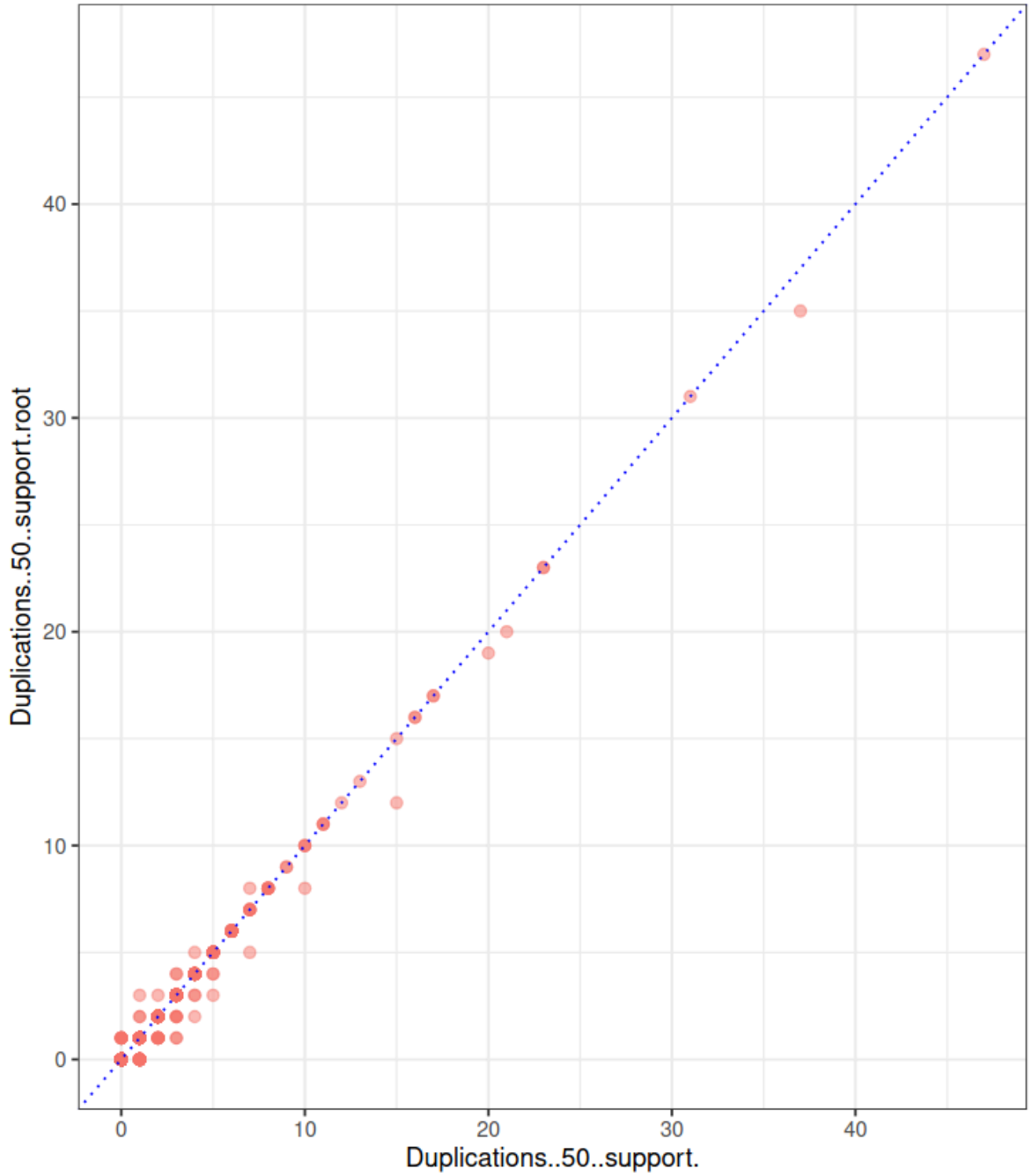


Figure 5. Comparaison de la distribution des duplications d'orthogroupes chez *Prochlorococcus* calculée par OrthoFinder avec (axe des ordonnées) et sans (axe des abscisses) enracinement de l'arbre d'espèces. Chaque point correspond à un orthogroupe pour lequel des événements de duplication ont été identifiés, l'intensité du point reflétant le nombre relatif de duplications observées. La droite en pointillés représente la diagonale d'égalité, où les valeurs coïncident parfaitement entre les deux analyses.
La forte corrélation observée le long de cette diagonale indique que l'enracinement de l'arbre n'a qu'un effet marginal sur la détection des duplications, suggérant une stabilité des résultats d'**OrthoFinder** indépendamment de la position de la racine.

Le seuil de **support ≥ 50** appliqué dans cette analyse signifie qu'un événement de duplication n'est retenu que s'il est observé dans au moins 50 % des espèces de part et d'autre des branches concernées, assurant ainsi la robustesse et la fiabilité de l'interprétation.

Analyse du pan-génome

À partir des orthogroupes identifiés, une **analyse pan-génomique** a été menée afin d'évaluer la diversité génétique du collectif *Prochlorococcus*. Cette approche consiste à distinguer trois composantes principales : le **génome cœur**, commun à toutes les souches ; le **génome accessoire**, partagé par un sous-ensemble de souches ; et les **gènes spécifiques**, présents dans une seule souche.

Les données issues d'OrthoFinder ont été analysées sous **R**, à l'aide des bibliothèques *ggplot2* et *matrixStats*, ainsi que du script *panplots.R*, permettant de générer les **courbes d'accumulation** du pan-génome et du génome cœur. Ces courbes décrivent l'évolution du nombre total de gènes observés en fonction du nombre de génomes ajoutés à l'analyse, selon la loi de **Heaps**. Cette loi permet de déterminer si le pan-génome est **ouvert** (diversité génétique croissante à mesure que de nouveaux génomes sont ajoutés) ou **fermé** (diversité limitée et stable).

La loi de Heaps, revue par *Tettelin et al.(2008),* modélise la relation entre le nombre total de gènes uniques observés (n), et le nombre de génomes analysés (N), avec k une constante empirique liée au jeu de données (https://hal.science/hal-04822722v1/file/Bioinfomics.pdf).

$$n = k \times N^{\gamma}$$

$0 < \gamma < 1$: pan-génome **ouvert**
$\gamma < 0$: pan-génome **fermé**

Les résultats ont montré que le pan-génome de *Prochlorococcus* est ouvert (Fig.6), témoignant d'une forte plasticité génétique et d'une capacité d'adaptation élevée à différents environnements marins. L'intérêt de cette analyse réside dans sa capacité à relier la diversité génomique aux stratégies écologiques des organismes. Toutefois, cette approche repose sur des hypothèses simplificatrices et reste sensible à la qualité des annotations et à la représentativité de l'échantillon analysé.

Sélection des familles

La sélection des familles de gènes conservés a été effectuée par un autre membre du groupe de travail. Cette étape visait à isoler un **sous-ensemble de gènes orthologues uniques** présents dans toutes les souches, formant ainsi la base du **génome cœur** utilisé pour les analyses phylogénétiques.

Les alignements multiples de séquences ont été réalisés à l'aide de **T-Coffee**, un logiciel permettant d'évaluer la cohérence des alignements à l'aide du **Transitive Consistency Score (TCS)**. Seuls les alignements présentant un score élevé ont été conservés, garantissant une qualité optimale pour la reconstruction phylogénétique. Les alignements ont ensuite été filtrés pour éliminer les gènes présentant des duplications (paralogues) ou des séquences incomplètes, afin de ne retenir que les familles orthologues strictes.

L'intérêt de cette approche réside dans la **fiabilité des alignements retenus**, essentiels pour obtenir des arbres espèces cohérents. Ses limites proviennent du fait qu'elle repose sur une sélection restreinte du génome cœur, qui ne reflète pas nécessairement toute la diversité fonctionnelle du groupe étudié.

Construction des arbres espèces

La dernière étape du projet a consisté à **inférer les arbres phylogénétiques** décrivant les relations évolutives entre les souches de *Prochlorococcus* et de *Synechococcus*. Ces arbres ont été obtenus à partir des alignements multiples sélectionnés précédemment, selon plusieurs approches complémentaires.

Les alignements nucléotidiques concaténés ont été utilisés pour inférer un **arbre d'espèces par maximum de vraisemblance** à l'aide du logiciel **IQ-TREE**. Plusieurs modèles de substitution ont été testés : un modèle à codons (**GY+F+R4**) (Fig.7), un modèle déterminé automatiquement par **ModelFinder (MFP)**, et des modèles protéiques (**WAG, LG, JTT**) (Fig.8) enrichis de variantes mixtes (**LG4M, LG4X**) pour modéliser l'hétérogénéité des taux d'évolution entre sites. Les valeurs de bootstrap et de test d'approximation de vraisemblance (aLRT) ont été utilisées pour estimer la robustesse des branches.

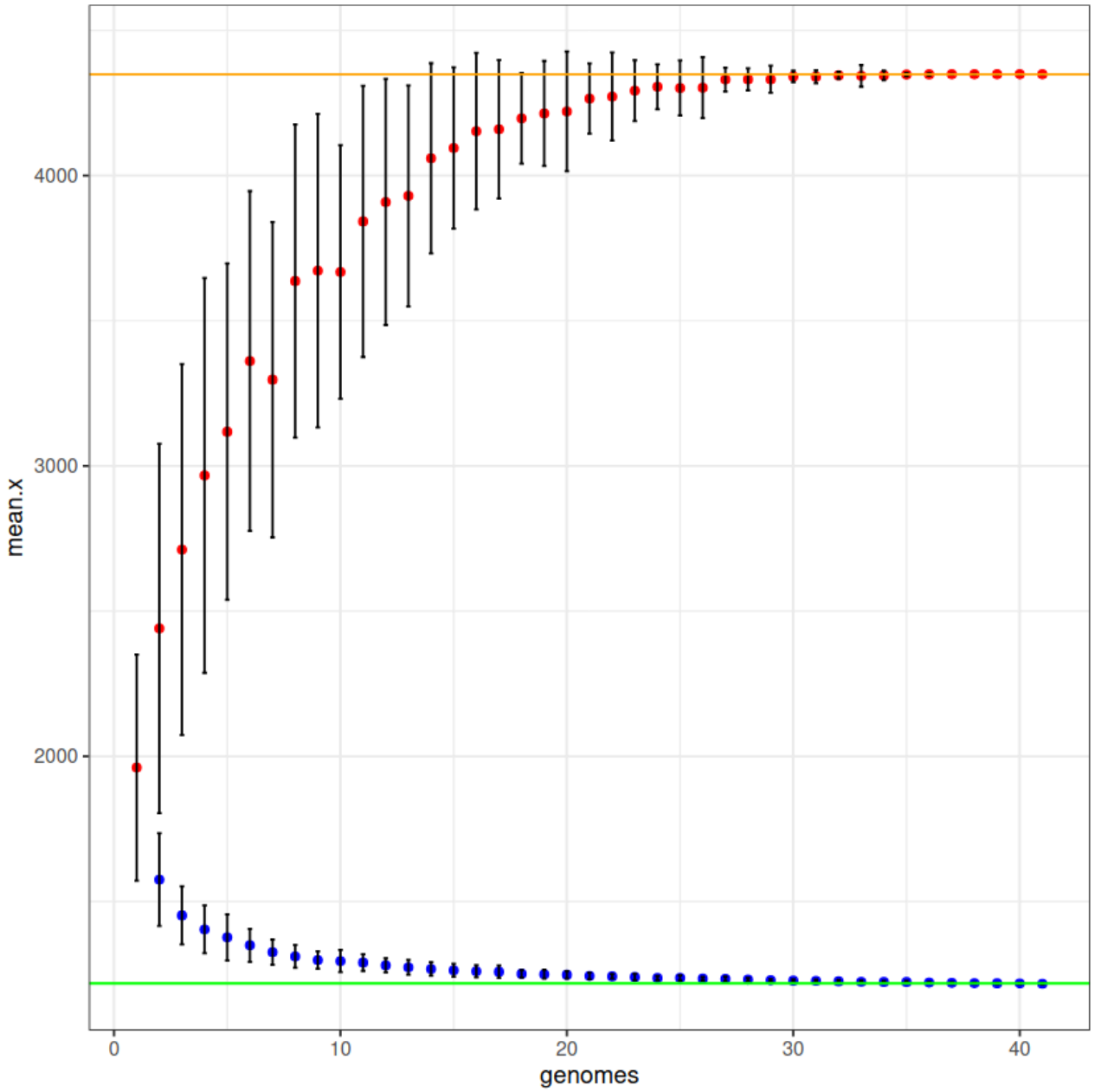


Figure 6a. Courbes d'accumulation du pan-génome (points rouges) et du génome cœur (points bleus) de *Prochlorococcus*, calculées à partir de la matrice présence/absence des orthogroupes (OrthoFinder). Pour chaque taille d'échantillon (abscisse : nombre de génomes), les valeurs moyennes (ordonnée) résultent de permutations aléatoires des génomes (100 itérations). Les barres noires représentent ± 1 écart-type. Les lignes horizontales indiquent les plateaux estimés sur l'ensemble des 40 génomes (en haut : pan-génome = **4349** gènes ; en bas : génome cœur = **1218** gènes). Le plateau atteint par la courbe rouge présente un **pan-génome fermé**, tandis que la courbe bleue se stabilise, indiquant un **cœur** relativement constant.

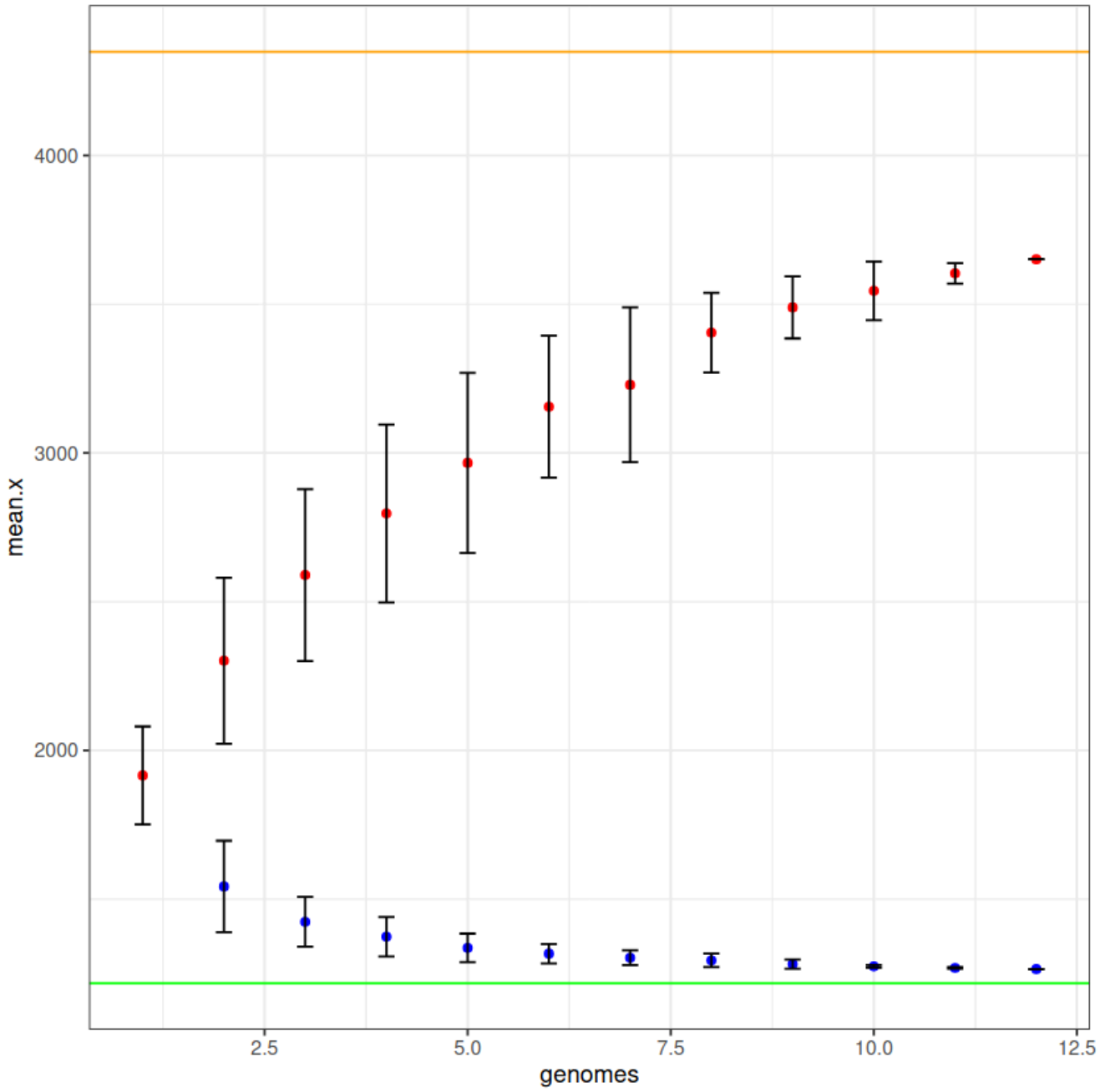


Figure 6b. Courbes d'accumulation du pan-génome (points rouges) et du génome cœur (points bleus) de *Prochlorococcus*, calculées à partir de la matrice présence/absence des orthogroupes (OrthoFinder), restreinte au sous-échantillon de Zhang (12 génomes). Les moyennes et écarts-types sont obtenus par permutations (100 itérations). Les lignes horizontales haute (= **4349**) et basse (= **1218**) reprennent les **plateaux de référence** estimés sur les 40 génomes pour faciliter la comparaison. La tendance observée est : **augmentation continue** du pan-génome et **stabilisation** du génome cœur, ce qui est cohérent avec une **forte plasticité génétique** du collectif et un noyau fonctionnel conservé.

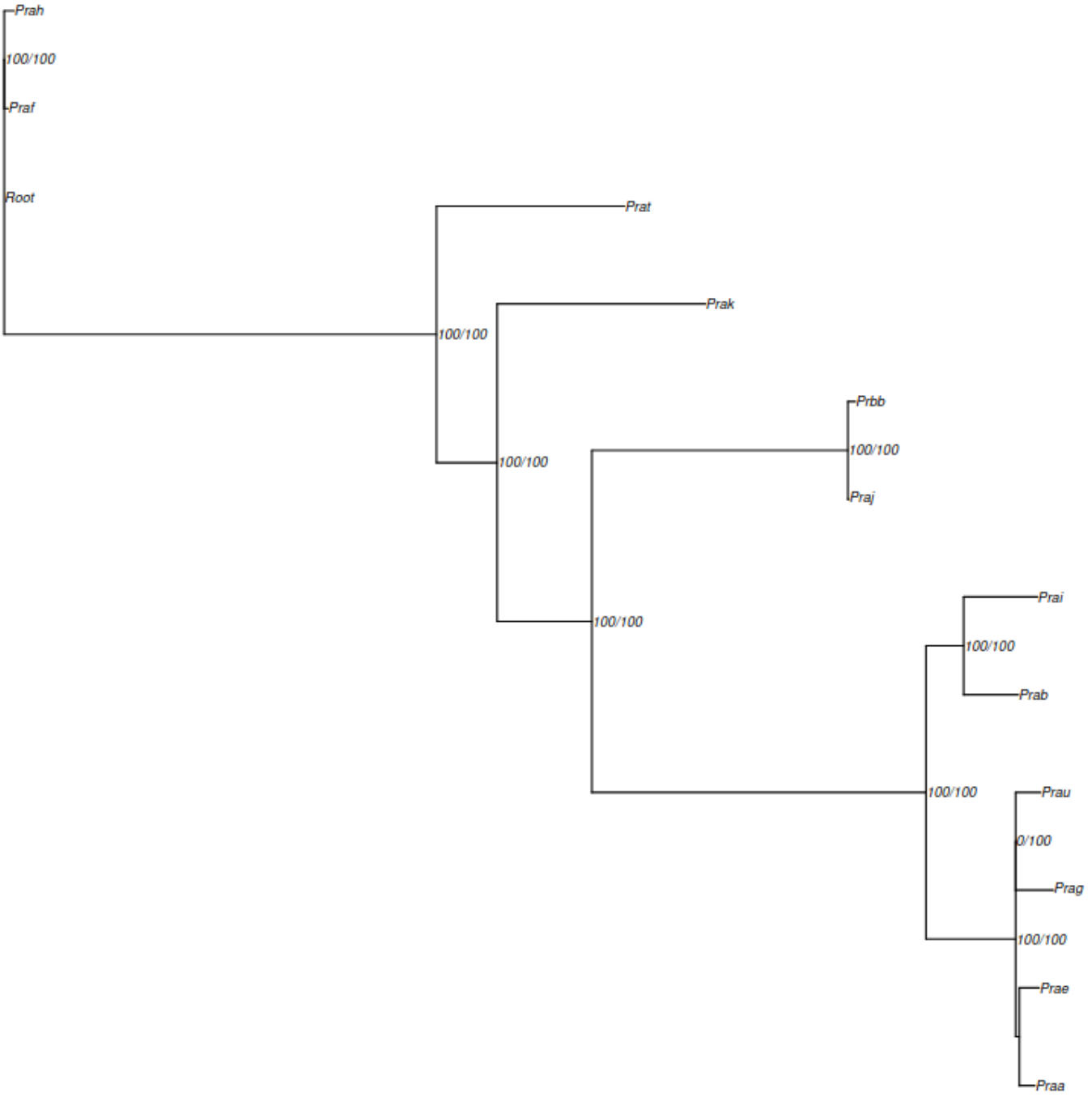


Figure 7. Arbre phylogénétique des souches de *Prochlorococcus* obtenu à partir du super-alignement codonique sous IQ-TREE avec le modèle GY+F+R4, puis enraciné sur les souches profondes *Prah* et *Praf*. L'arbre met en évidence une structuration claire entre les écotypes de haute lumière (HL) et de basse lumière (LL).

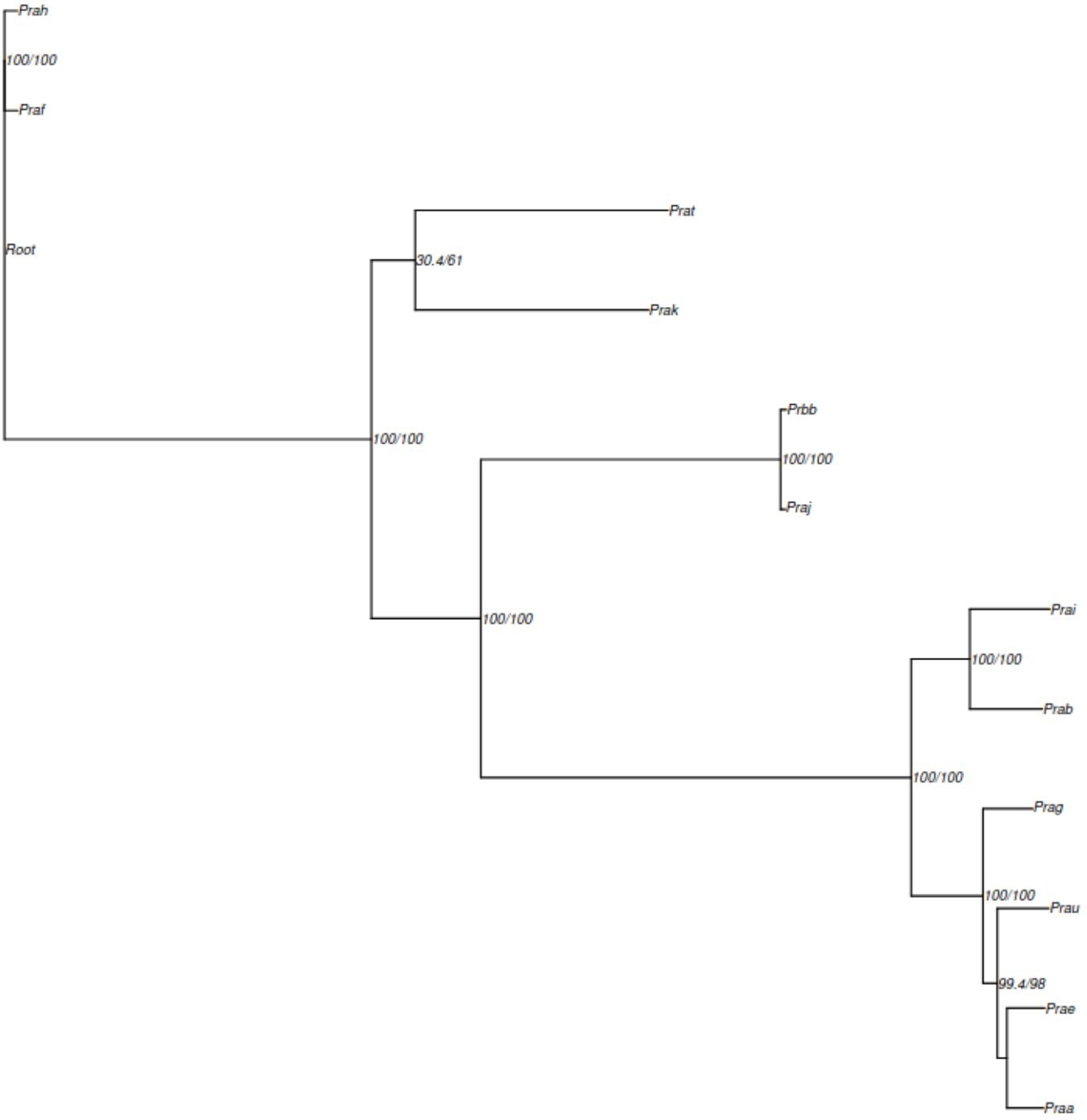


Figure 8. Arbre phylogénétique des souches de *Prochlorococcus* obtenu à partir de l'alignement des séquences protéique, inféré par maximum de vraisemblance sous IQ-TREE. L'arbre est reraciné sur les souches profondes *Prah* et *Praf*, révélant une séparation nette entre les écotypes de haute lumière (HL) et de basse lumière (LL).

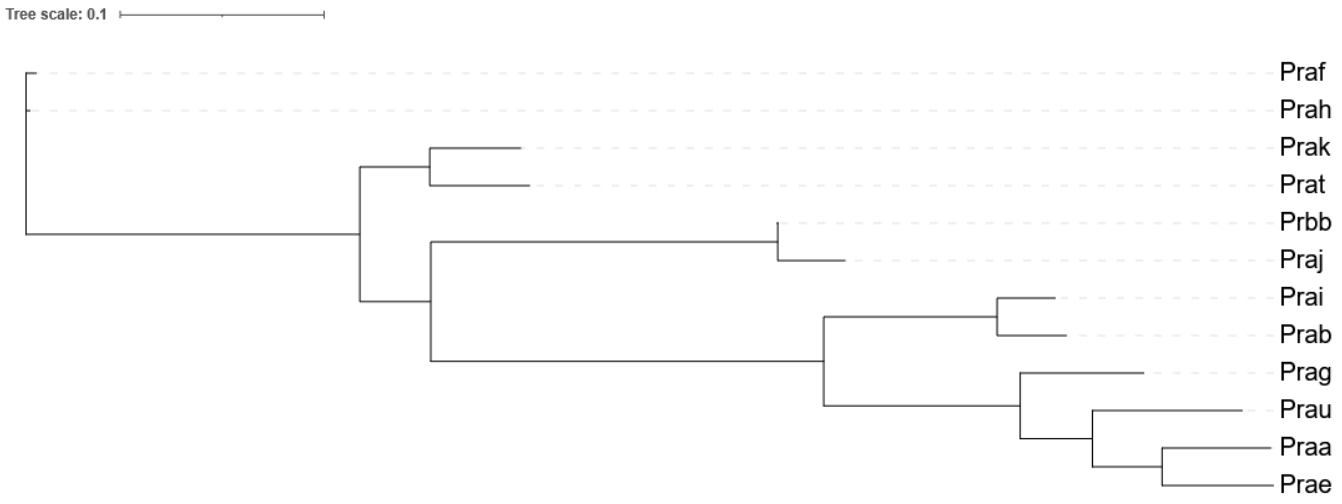


Figure 9. Arbre phylogénétique des souches de *Prochlorococcus* obtenu à partir du super-arbre MRP (Matrix Representation Parsimony) construit à partir des arbres individuels de gènes. Chaque topologie d'arbre de gène a été traduite en une matrice binaire, puis combinée pour générer un arbre consensus reflétant les relations phylogénétiques globales entre souches. La figure met en évidence la cohérence des regroupements majeurs observés dans les autres approches, avec une séparation nette entre les écotypes de haute lumière (HL) et de basse lumière (LL).

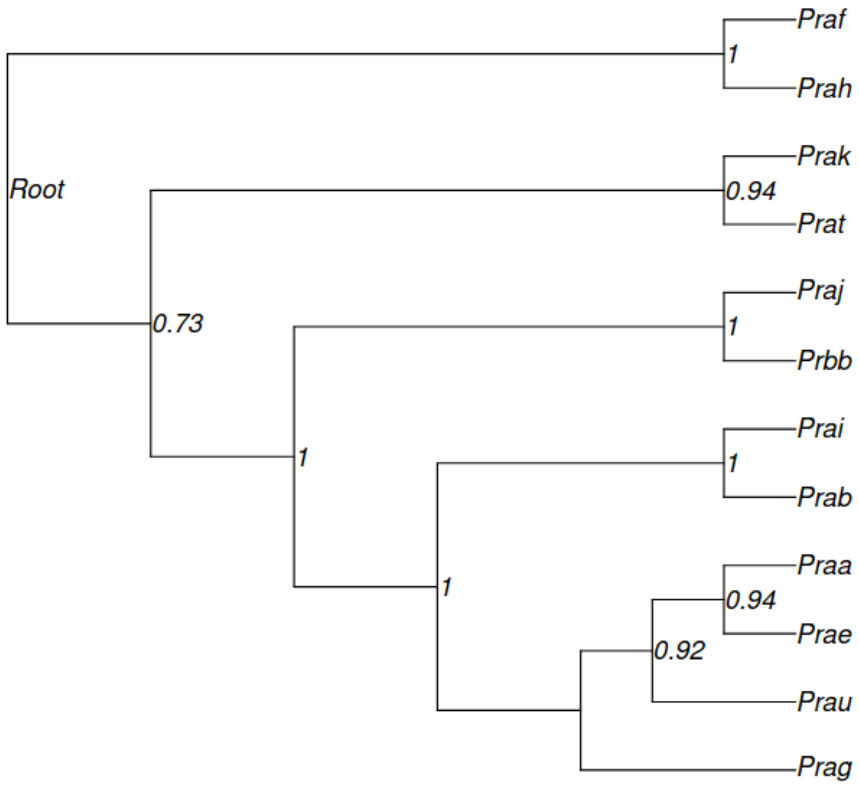


Figure 10. Arbre d'espèces enraciné obtenu avec la méthode ASTRAL à partir des arbres de gènes individuels. Cette approche de super-arbre révèle une topologie cohérente avec les modèles précédents, distinguant clairement les clades de *Prochlorococcus* adaptés aux fortes et faibles intensités lumineuses.

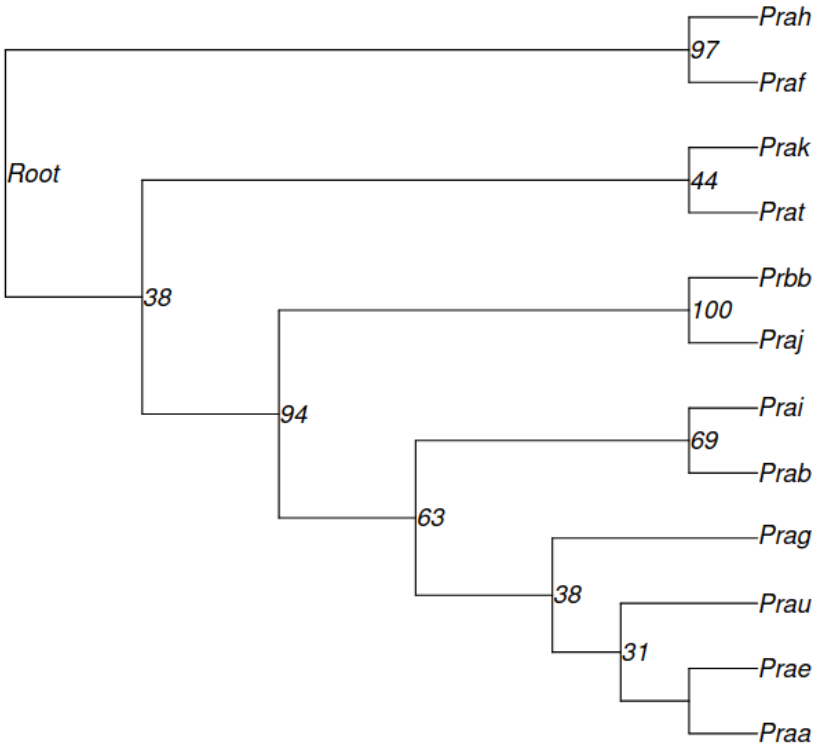


Figure 11. Arbre phylogénétique obtenu par la méthode du super-arbre consensus des souches de *Prochlorococcus*, puis enraciné sur *Prah* et *Praf*. La topologie résultante synthétise les relations les plus robustes entre écotypes et confirme la distinction nette entre les clades de haute et basse lumière.

Parallèlement, des **super-arbres** ont été construits à partir des arbres individuels de gènes protéiques, selon trois méthodes distinctes : (1) la **méthode MRP (Matrix Representation Parsimony)** (Fig.9), qui fusionne les arbres gènes en une matrice binaire pour en déduire un arbre consensus ; (2) la **méthode ASTRAL** (Fig.10), qui infère un arbre espèces à partir des topologies locales de quartets les mieux soutenues ; et (3) l'arbre consensus majoritaire (Fig.11), calculé avec **IQ-TREE** et **R (ape, phytools)**.

La comparaison entre les arbres a révélé une différence topologique majeure liée au choix des modèles et des méthodes d'enracinement. La méthode de construction d'arbre basée sur l'alignement des séquences nucléotidiques est la seule dont l'arbre final est distinct des autres méthodes. Les distances de **Robinson-Foulds** (Fig.12) ont été calculées pour quantifier ces écarts. L'intérêt de ces approches multiples est de tester la robustesse des inférences phylogénétiques, tandis que leurs limites résident dans la sensibilité aux erreurs d'alignement et à la qualité du jeu de données.

```
5 5
Tree0       0 4 4 4 4
Tree1       4 0 0 0 0
Tree2       4 0 0 0 0
Tree3       4 0 0 0 0
Tree4       4 0 0 0 0
```

Figure 12. Matrice de distance de Robinson-Foulds. Les arbres ont été ajoutés lors du calcul dans l'ordre suivant : (0) arbre sur les codons (modèle GYFR4), (1) arbre sur les protéines, (2) super-arbre consensus, (3) super-arbre MRP et (4) super-arbre ASTRAL. La méthode la moins robuste pour notre jeu de données est celle basée sur les séquences nucléotidiques.

Conclusion et Discussion

Discussion

Ce projet a mis en place une chaîne d'analyse phylogénomique complète — de la constitution du jeu de données à l'inférence d'arbres espèces — appliquée au collectif *Prochlorococcus* (avec *Synechococcus* comme groupe externe). Plusieurs enseignements se dégagent.

Qualité et composition du jeu de données. Le nettoyage initial (détection des génomes de phages, contrôle complétude/contamination via CheckM) s'est révélé déterminant : il a limité les biais en aval (fausses absences, faux orthologues) et amélioré la stabilité des arbres. La distinction entre (i) un jeu de données complet issu de RefSeq et (ii) un sous-échantillon réduit (Zhang) a permis d'équilibrer couverture biologique et coût de calcul. Cette stratégie a toutefois un revers : le sous-échantillon, plus maniable, peut sous-représenter certaines lignées et aplanir des signaux évolutifs rares.

Annotations et orthogroupes. Le couplage Prokka avec eggNOG-mapper a fourni des annotations homogènes et exploitables, mais demeure sensible à la couverture des bases et conduit encore à un nombre non négligeable de « protéines hypothétiques ». L'inférence d'orthogroupes par OrthoFinder, robuste et phylogénétiquement informée, a servi de socle aux analyses de pan-génome et aux reconstructions d'arbres. On note une bonne cohérence entre matrices présence/absence, heatmaps de chevauchement et statistiques de duplications. La comparaison « duplications avec vs sans enracinement » montre une forte corrélation, suggérant que la position de la racine influe peu sur la détection des duplications au seuil de support retenu (≥ 0,5). Les cas discordants pourraient refléter des duplications récentes, des transferts horizontaux ou des incertitudes locales de topologie.

Pan-génome : ouvert, fermé… ou entre les deux ? Sur l'ensemble de 40 génomes, la courbe d'accumulation tend vers un palier (4349 gènes), laissant penser que, pour **cet** échantillonnage, la diversité supplémentaire capturable se raréfie (comportement proche d'un pan-génome « quasi-fermé »). À l'inverse, sur le sous-échantillon de Zhang (12 génomes), la courbe continue de croître, ce qui est compatible avec un pan-génome **ouvert**. Cette apparente contradiction s'explique par (i) la taille et la composition de l'échantillon, (ii) la redondance entre souches proches, et (iii) le fait que la loi de Heaps dépend fortement de la diversité réellement couverte. En pratique, le collectif *Prochlorococcus* reste réputé très plastique ; notre analyse indique surtout que la saturation observée à 40 génomes reflète les limites de **ce** jeu plutôt qu'un caractère intrinsèque « fermé » du pan-génome.

Arbres espèces et cohérence inter-méthodes. Les approches super-matrice (IQ-TREE, modèles codoniques et protéiques) et super-arbres (MRP, ASTRAL, consensus) convergent globalement : séparation nette des écotypes haute lumière (HL) / basse lumière (LL), stabilité des clades majeurs, et topologies proches malgré des modèles différents (GY+F+R4, MFP, LG/LG4M/LG4X). Les quelques divergences tiennent à des choix de modèles, à la présence résiduelle de paralogues mal filtrés, et aux limites des alignements multiples. L'usage de mesures complémentaires (bootstrap ultrarapide, aLRT, distances de Robinson–Foulds) a permis d'objectiver la robustesse des nœuds, mais un chiffrage par **concordance factors** (gCF/sCF) ou par **quartets** renforcerait encore l'évaluation.

Limites principales. (i) Biais d'échantillonnage (sur-représentation de certaines lignées HL/LL), (ii) incertitudes d'annotation/fonction (bases incomplètes), (iii) paralogie et HGT pouvant perturber l'hypothèse d'orthologie stricte, (iv) sensibilité aux choix d'alignement et de modèle (hétérotachie, composition biaisée), et (v) non-prise en compte explicite de la recombinaison. Ces limites n'invalident pas les tendances observées mais appellent des analyses complémentaires.

Conclusion

Nous avons mis en œuvre un pipeline reproductible d'annotation, d'orthologie, de pan-génomique et de phylogénomique appliqué au collectif *Prochlorococcus*. Après contrôle de qualité et harmonisation, les analyses montrent : (1) une structuration phylogénétique robuste, cohérente entre approches super-matrice et super-arbres, avec la séparation HL/LL clairement résolue ; (2) un génome cœur stable et un répertoire accessoire substantiel ; (3) un comportement d'accumulation indiquant un pan-génome encore expansible lorsque l'on échantillonne des

lignées plus diversifiées, malgré un palier apparent sur l'échantillon complet utilisé ; (4) une détection des duplications globalement stable quel que soit l'enracinement.

Au-delà de ces résultats, le projet valide la pertinence d'une démarche graduelle (du jeu complet vers un sous-échantillon représentatif) pour concilier **portée biologique** et **coût computationnel**, tout en conservant des conclusions solides.

Perspectives

1. Étendre et équilibrer l'échantillonnage (écotypes, régions, saisons) ; estimer rigoureusement l'exposant γ de la loi de Heaps (avec intervalles de confiance) et répéter les courbes avec **rarefaction**/ré-échantillonnage contrôlé.

2. Raffiner les délimitations taxonomiques avec **ANI/AAI**, **GTDB-Tk** et **dDDH**, et comparer aux propositions récentes de genres.

3. Renforcer l'évaluation des arbres par **concordance factors (gCF/sCF)**, supports **quartets** et tests d'**incongruence** (p.ex. **tree certainty**), tout en testant des modèles plus riches (partitionnement par gène/codon, **PMSF**, modèles site-hétérogènes ; à terme **CAT-GTR**).

4. Prendre en compte la **recombinaison** (p.ex. **ClonalFrameML**, **Gubbins**) et distinguer plus finement orthologues/paralogues récents.

5. Relier l'évolution des familles de gènes à des **fonctions écologiques** (voies phototrophes, utilisation de nutriments, stress) via des enrichissements GO/KEGG et analyses de **syntenie**.

En synthèse, ce travail fournit une base robuste et reproductible pour l'étude évolutive de *Prochlorococcus* ; il met en évidence une architecture génomique à la fois conservée (cœur) et hautement flexible (accessoire), motrice des adaptations écologiques qui font le succès planétaire de ce micro-organisme.

Références

- *Zhang et al., 2021 Snowball Earth, population bottleneck and Prochlorococcus evolution.*
- *Tettelin, Hervé et al. "Comparative genomics: the bacterial pan-genome." Current opinion in microbiology vol. 11,5 (2008): 472-7. doi:10.1016/j.mib.2008.09.006*
- *Kettler et al., PLoS Genet. 2007 Dec;3(12):e231 Patterns and implications of gene gain and loss in the evolution of ''Prochlorococcus".*
- *Sun and Blanchard, 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*
- *Yan et al., Appl Environ Microbiol. 2018 Genome rearrangement shapes ''Prochlorococcus" ecological adaptation.*
- *Yan et al., mBio 2022 Diverse Subclade Differentiation Attributed to the Ubiquity of Prochlorococcus High-Light-Adapted Clade II*
- *Biller et al., Nat. Rev. Microbiol. 2015 13(1) 13-27 ''Prochlorococcus": the structure and function of collective diversity.*
- *Partensky and Laurence Garczarek Annual Review of Marine Science 2010 Prochlorococcus: Advantages and Limits of Minimalism.*
- *Tschoeke et al., 2020 Unlocking the Genomic Taxonomy of the Prochlorococcus Collective.*
- *Yan et al., 2022 Diverse Subclade Differentiation Attributed to the Ubiquity of Prochlorococcus High-Light-Adapted Clade II.*
- *Ribalet et al., 2025 Future ocean warming may cause large reductions in Prochlorococcus biomass and productivity*
- *[Prochlorococcus] https://www.cell.com/current-biology/fulltext/S0960-9822(17)30213-0?code=cell-site*
- *[Cyanorak Information system] http://application.sb-roscoff.fr/cyanorak/welcome.html*

Reproductibilité

Toutes les analyses ont été réalisées sur la plateforme **GenoToul** (INRAE, Toulouse), dans un environnement Linux utilisant le gestionnaire de tâches **Slurm**.

Les fichiers d'entrée (séquences génomiques, annotations et métadonnées) sont récupérables à l'aide des scripts dans les répertoires de travail du projet, accompagnés des scripts Bash et R utilisés pour le traitement des données. Les jeux de données d'origine sont accessibles publiquement via la base **RefSeq (NCBI)** et via le dépôt associé à l'article de Zhang et al. (2021).

Les scripts personnalisés utilisés pour la préparation des données, la génération des commandes d'annotation et la visualisation des résultats sont disponibles sur le dépôt GitHub :

https://github.com/CamilleAstrid/fr.utoulouse.Phylogenomique_Prochlorococcus.git

Licence et propriété intellectuelle

Les données génomiques analysées proviennent exclusivement de sources publiques (NCBI RefSeq et publications scientifiques).

Les logiciels **Prokka**, **eggNOG-mapper**, **OrthoFinder**, **T-Coffee**, **IQ-TREE** et **ASTRAL** sont distribués sous licences libres ou académiques, et ont été utilisés sans modification de leur code source.

Les scripts Bash et R développés dans le cadre de ce projet sont des adaptations pédagogiques réalisées pour les besoins des travaux pratiques et appartiennent à leurs auteurs respectifs, sous la supervision des enseignants de l'unité d'enseignement de Phylogénomique (Yves Quentin et Gwennael Fichant). Les script pour l'automatisation de l'analyse a été adapté des scripts précédents par l'auteur de ce rapport.

Les résultats et figures produits sont destinés à un usage académique et non commercial, dans le cadre du Master de Bioinformatique et Biologie des Systèmes de l'Université de Toulouse.

Données supplémentaires

Les figures présentées dans cette section regroupent l'ensemble des **résultats complémentaires** obtenus au cours du projet, qui n'ont pas été inclus dans le corps principal du rapport afin d'en alléger la présentation.

Elles offrent un aperçu plus détaillé des analyses intermédiaires et des visualisations produites à chaque étape du pipeline de phylogénomique et de pan-génomique de *Prochlorococcus*.

Ces données comprennent notamment :

- des **analyses de qualité des génomes** ;
- des **représentations issues d'OrthoFinder, illustrant la composition du pan-génome, les duplications géniques et la répartition des orthogroupes** ;
- des **arbres phylogénétiques** obtenus selon différentes approches ;
- ainsi que plusieurs **figures comparatives** (cophyloplots, arbres enracinés, consensus et réseaux SplitsTree) permettant d'évaluer la cohérence des résultats entre méthodes.

Ces figures servent à **compléter et contextualiser les résultats principaux** présentés dans le rapport. Elles permettent de visualiser plus finement les tendances observées — telles que la variabilité génomique, la structuration phylogénétique du collectif *Prochlorococcus* et la robustesse des inférences réalisées — tout en documentant les analyses réalisées dans leur intégralité.

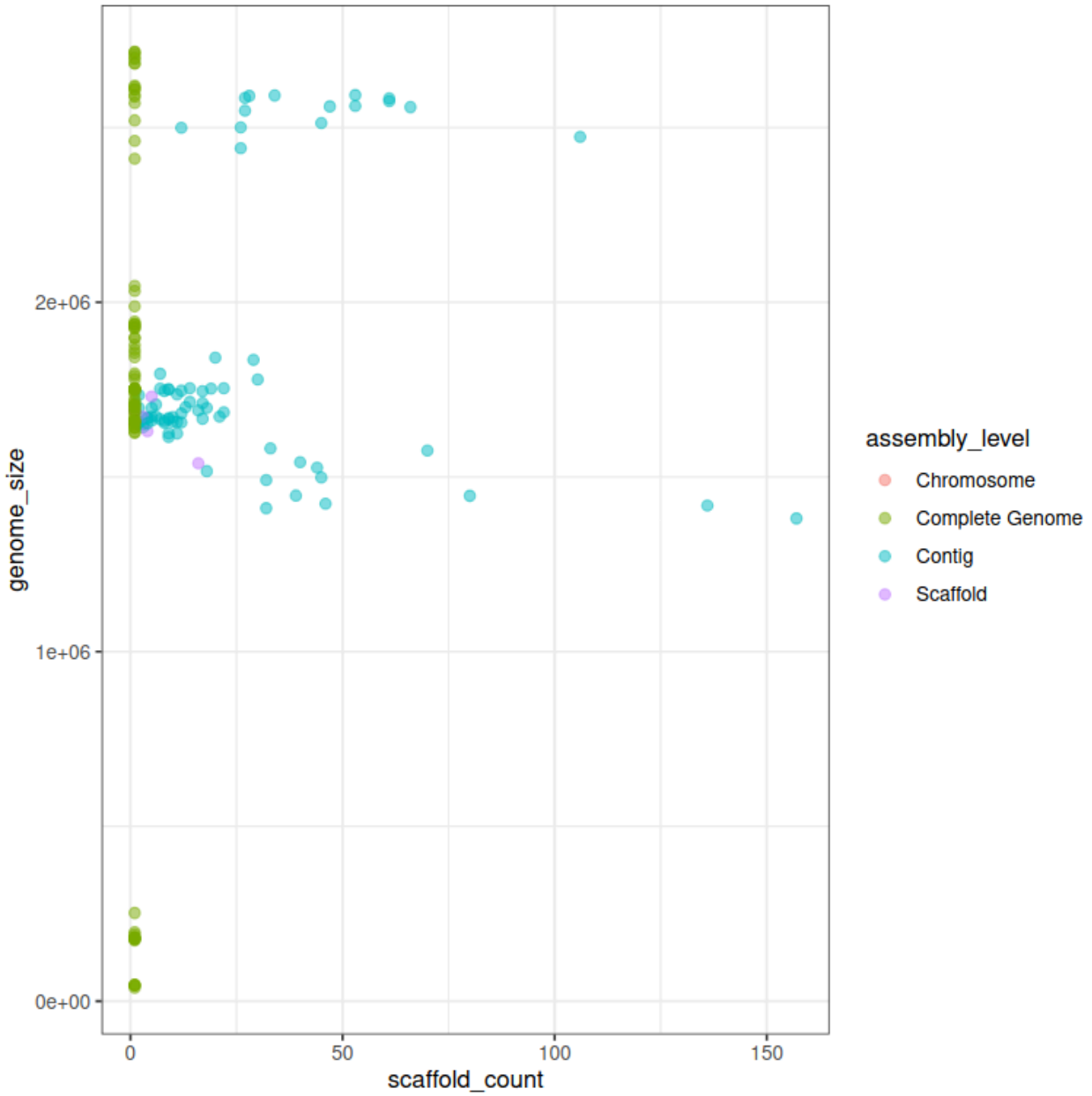


Figure 1 — Relation entre la taille totale du génome et le nombre de scaffolds pour les souches de *Prochlorococcus* issues de **RefSeq**. La figure met en évidence que les génomes les plus complets présentent généralement un nombre réduit de scaffolds, témoignant d'une meilleure qualité d'assemblage et d'une variabilité structurale modérée entre les souches.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/Genome_and_scaffold_size.pdf

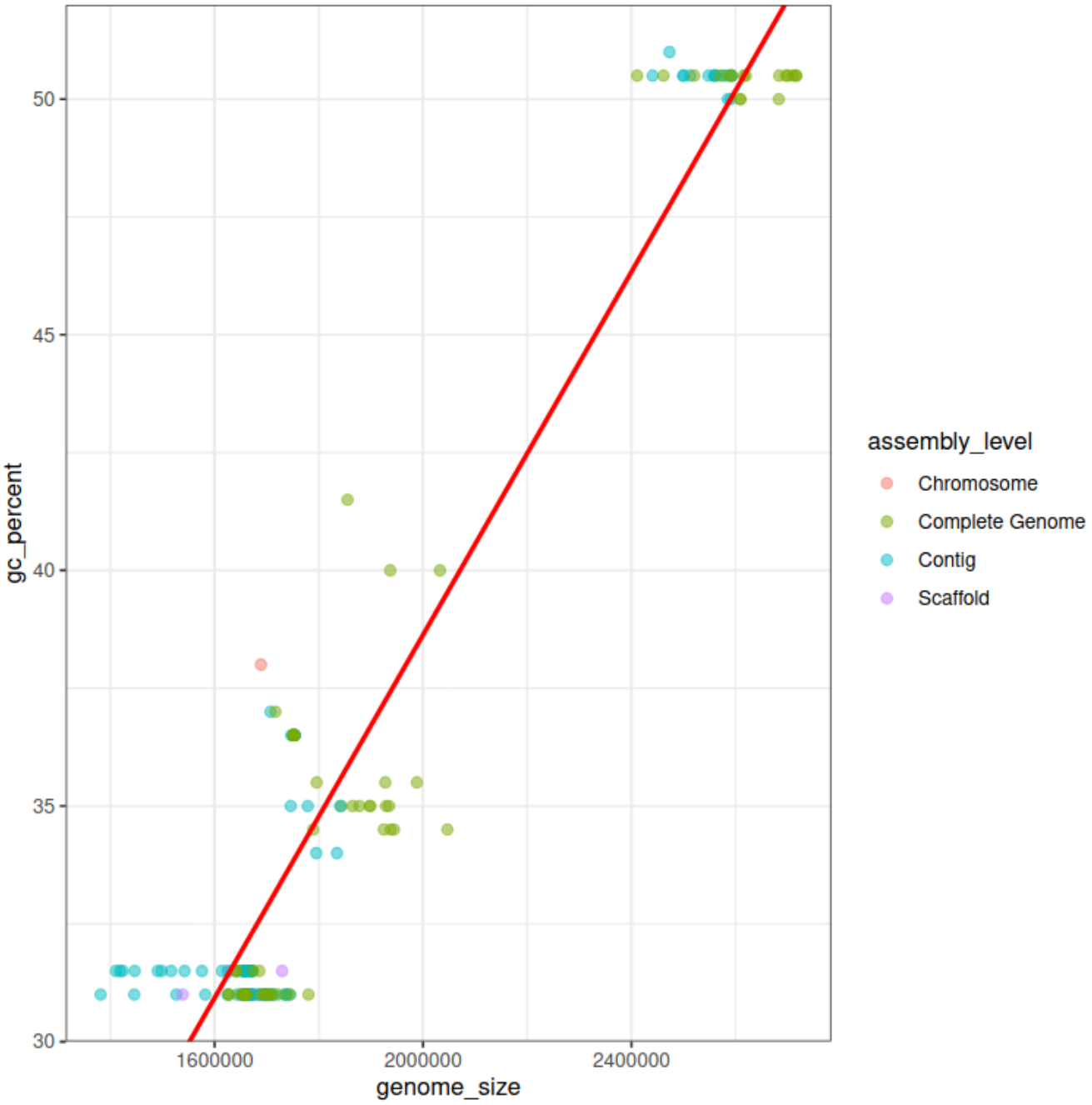


Figure 2 — Corrélation entre la taille du génome et le pourcentage de GC chez les souches de *Prochlorococcus* issues de **RefSeq**. La figure met en évidence une tendance générale à la réduction génomique accompagnée d'une diminution du contenu en GC, caractéristique des clades adaptés aux environnements oligotrophes et à forte luminosité.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/GC_percent_and_Genome_size.pdf

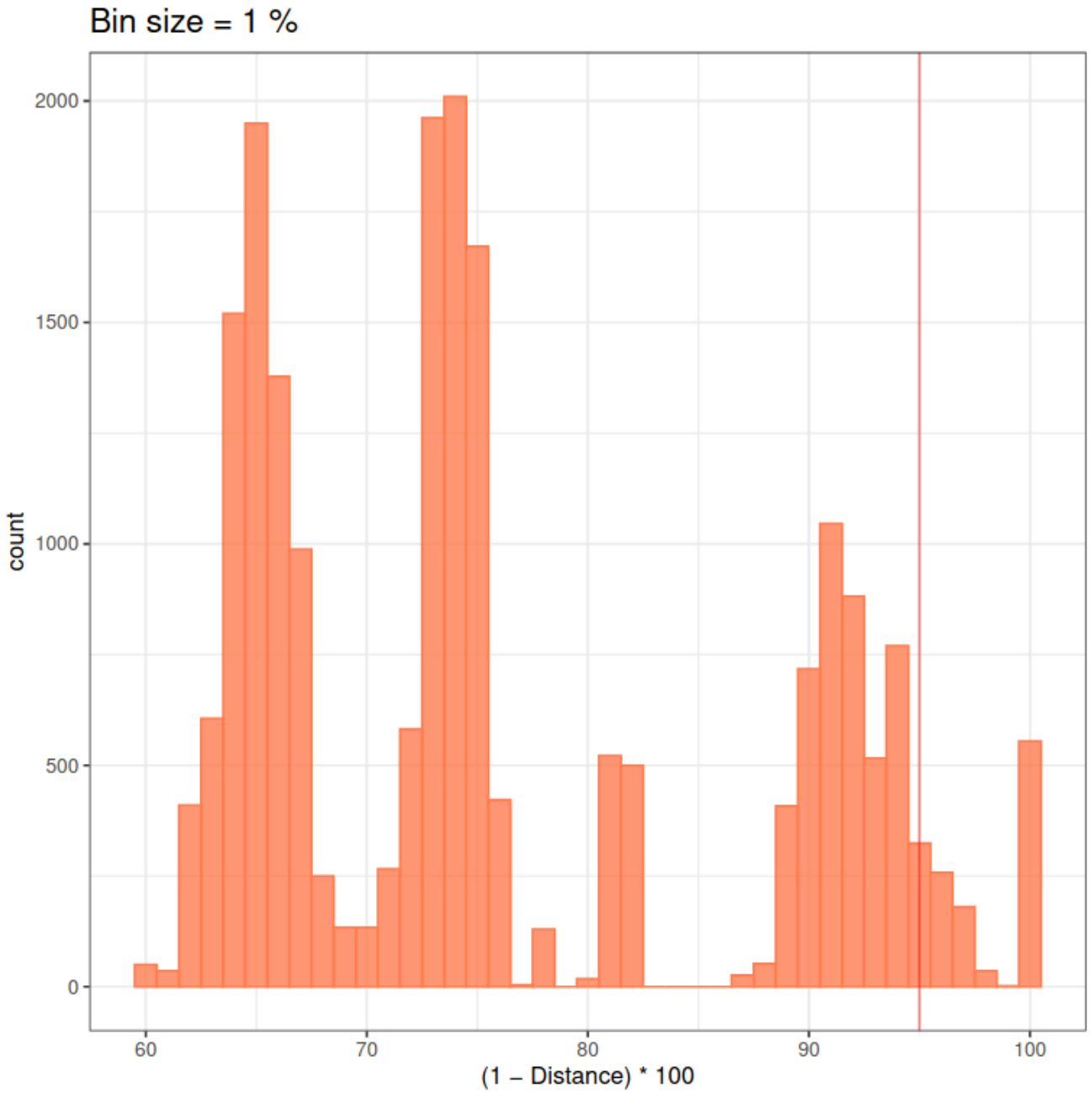


Figure 3 — Distribution des identités de paires génomiques calculées avec **Mash** pour les souches de *Prochlorococcus*. La figure montre la diversité génétique intra-genre, avec des regroupements correspondant aux principaux écotypes, reflétant la structure phylogénétique et écologique du collectif.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/MashProchlo/Mash_id_distribution.pdf

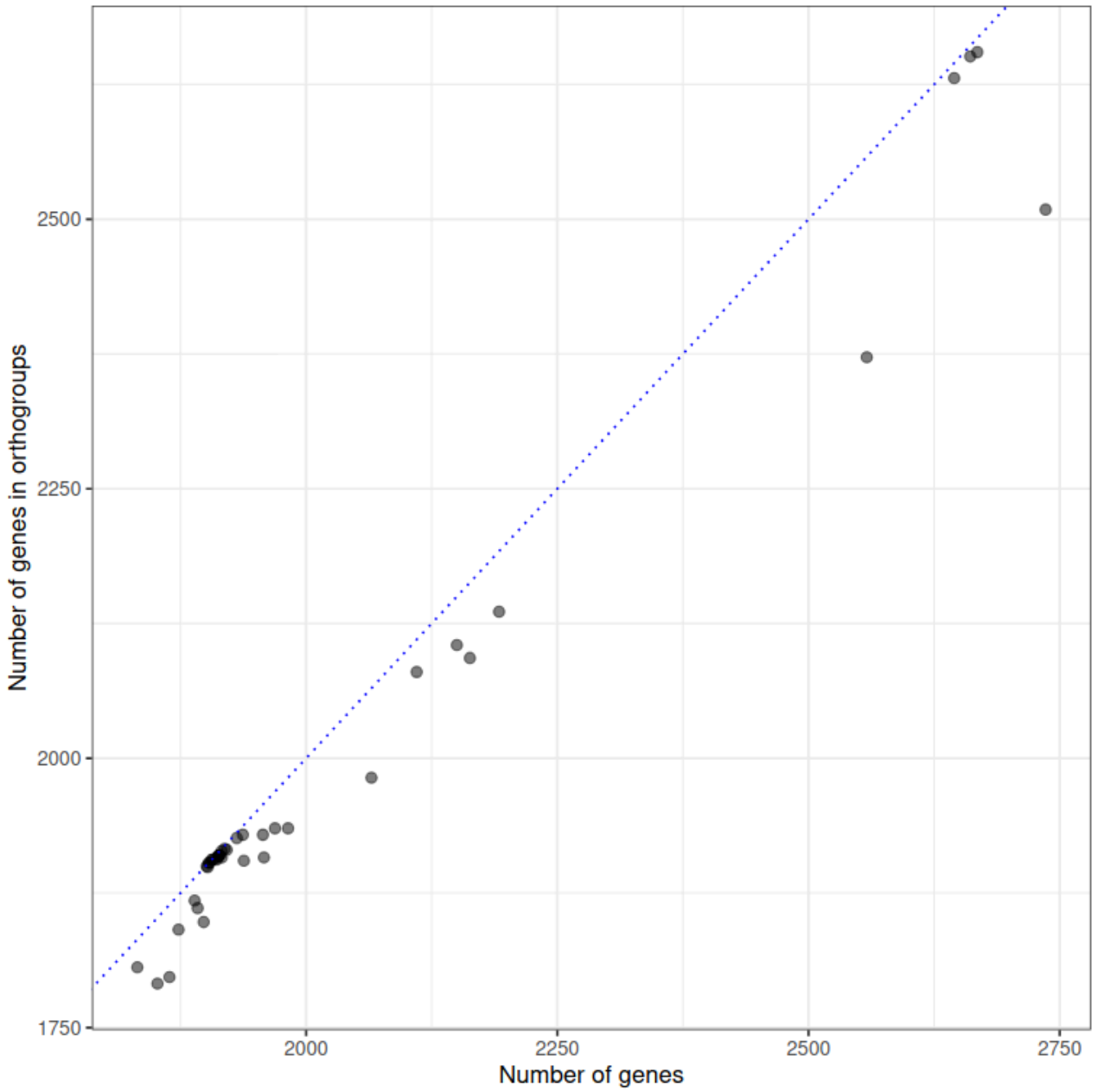


Figure 4 — Statistiques globales par espèce issues d'**OrthoFinder** pour les génomes de *Prochlorococcus*. La figure présente, pour chaque souche, le nombre total de gènes, d'orthogroupes partagés et de gènes spécifiques, mettant en évidence la variabilité génétique inter-souches et la structuration du collectif selon les écotypes.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/Comparative_Genomics_Statistics/Statistics_PerSpecies_global.pdf

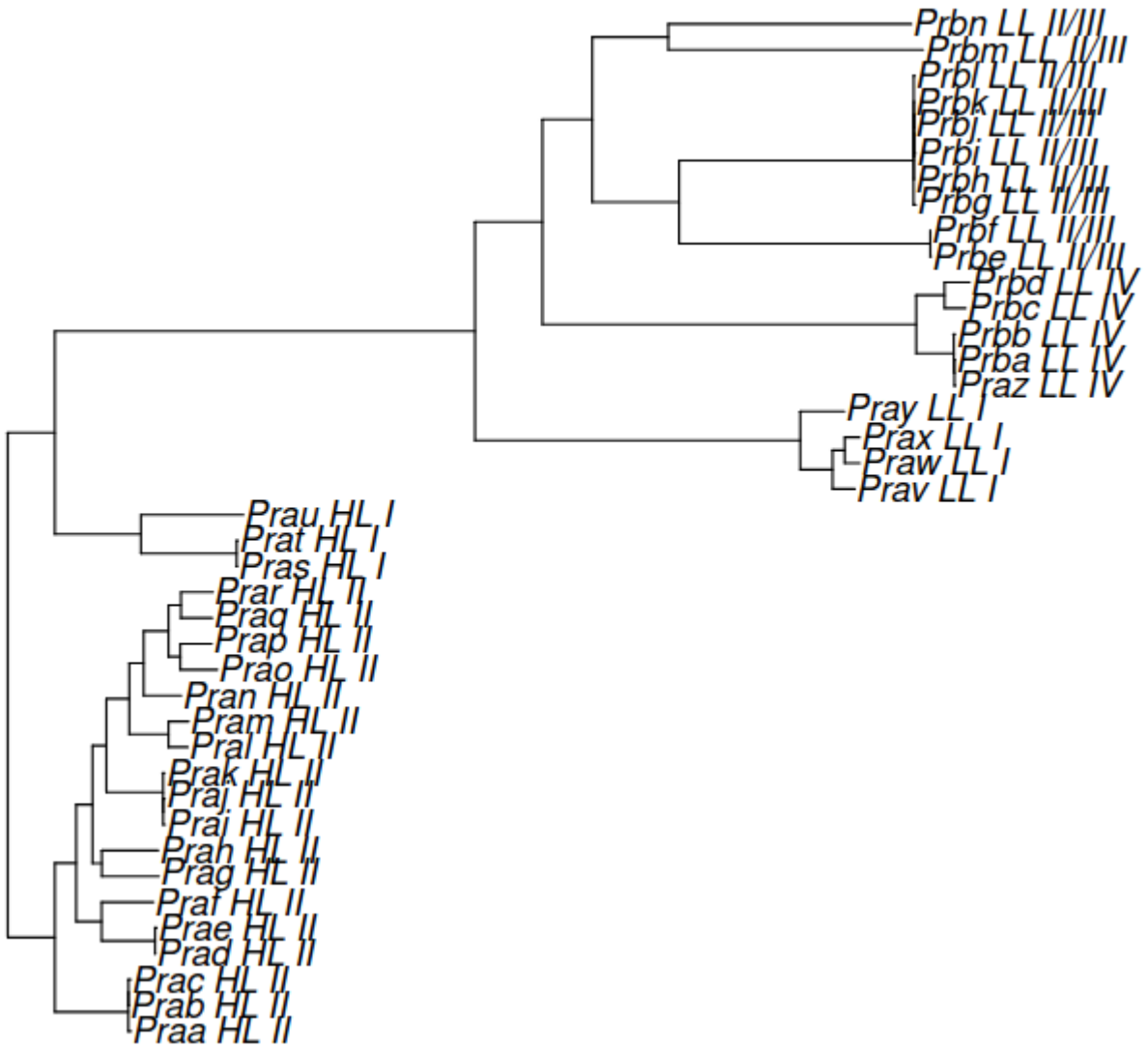


Figure 5 — Arbre d'espèces enraciné de *Prochlorococcus* obtenu avec **OrthoFinder** à partir des orthogroupes protéiques. La figure met en évidence la séparation nette entre les écotypes de haute lumière (HL) et de basse lumière (LL), reflétant la diversité évolutive et écologique du collectif.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/Species_Tree/SpeciesTree_rooted.pdf

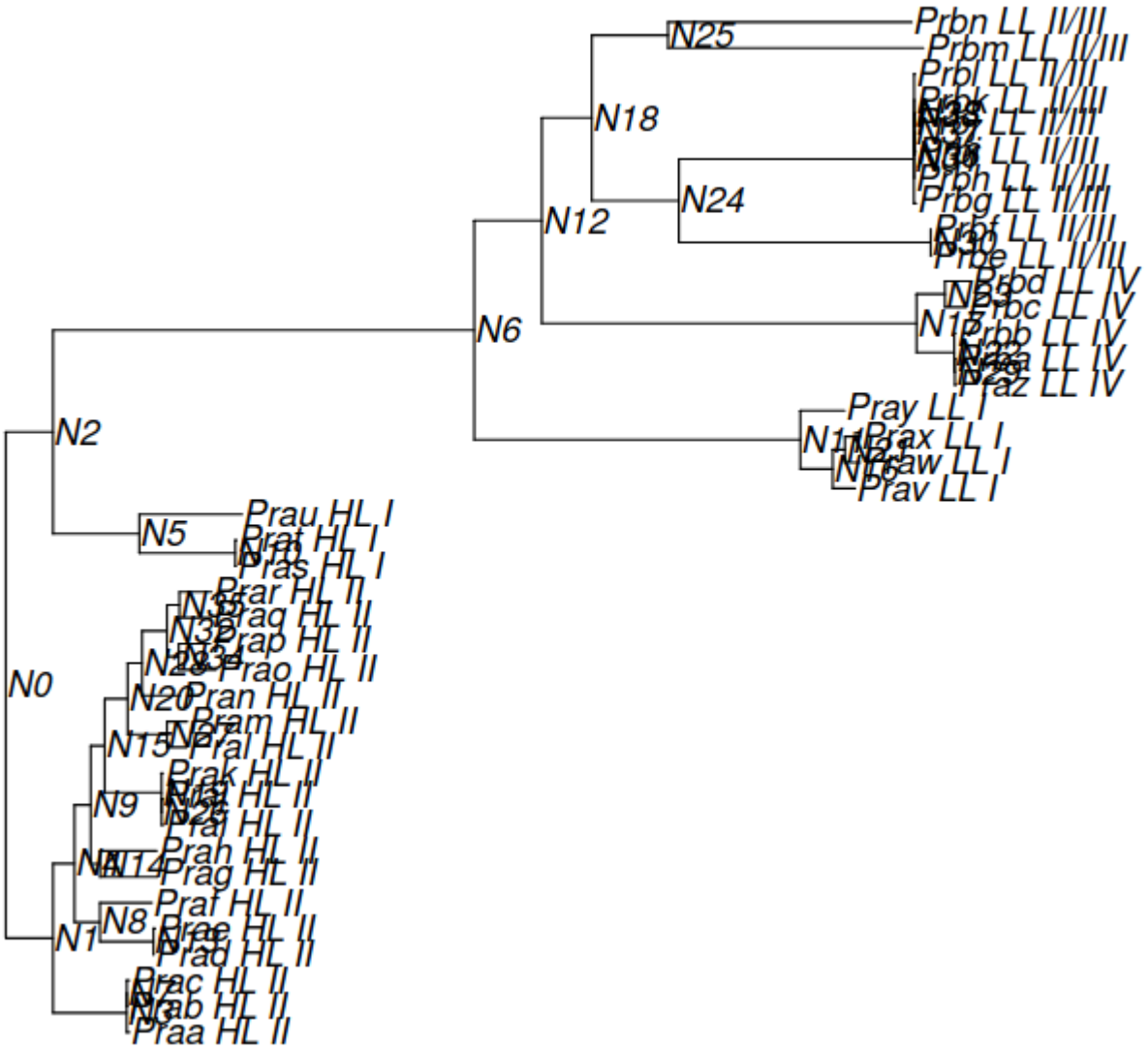


Figure 6 — Arbre d'espèces enraciné de *Prochlorococcus* généré par **OrthoFinder**, avec affichage des étiquettes de nœuds. Cette représentation détaille les points de divergence entre clades et met en évidence la structuration phylogénétique du collectif, distinguant clairement les écotypes de haute lumière (HL) et de basse lumière (LL).

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/Species_Tree/SpeciesTree_rooted_node_labels.pdf

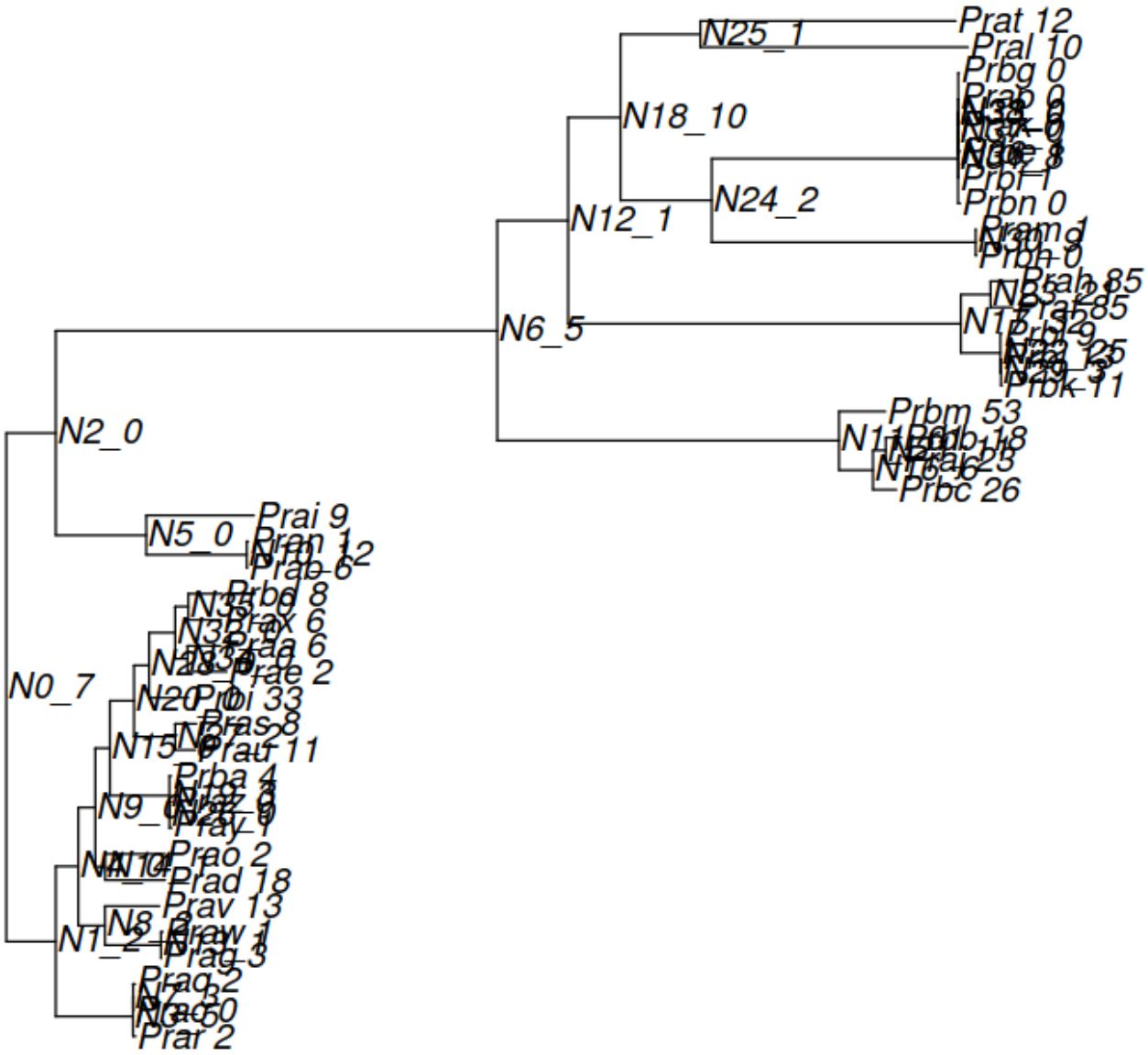


Figure 7 — Arbre d'espèces de *Prochlorococcus* annoté avec les événements de duplication génique identifiés par **OrthoFinder** (seuil de support ≥ 0,5). Les branches présentant des duplications fréquentes suggèrent des expansions géniques spécifiques à certains clades, probablement liées à des adaptations fonctionnelles et écologiques différenciées.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/Gene_Duplication_Events/SpeciesTree_Gene_Duplications_0.5_Support.pdf

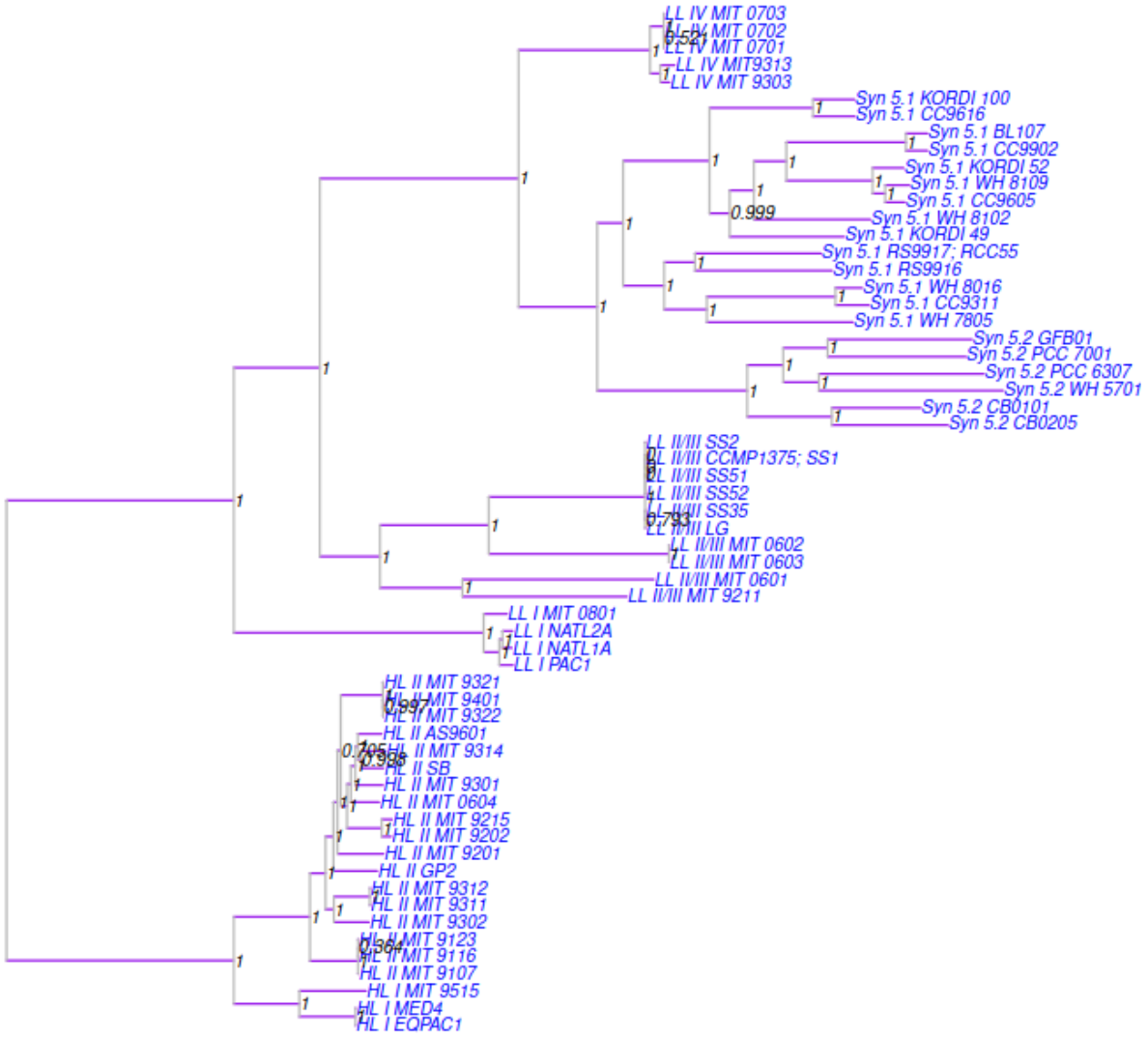


Figure 8 — Arbre d'espèces enraciné généré par **OrthoFinder** à partir des orthogroupes protéiques de *Prochlorococcus*. La figure illustre la structuration phylogénétique du collectif, avec une séparation claire entre les écotypes adaptés aux fortes intensités lumineuses (HL) et ceux des zones profondes à faible luminosité (LL).

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsa/Species_Tree/SpeciesTree_rooted.pdf

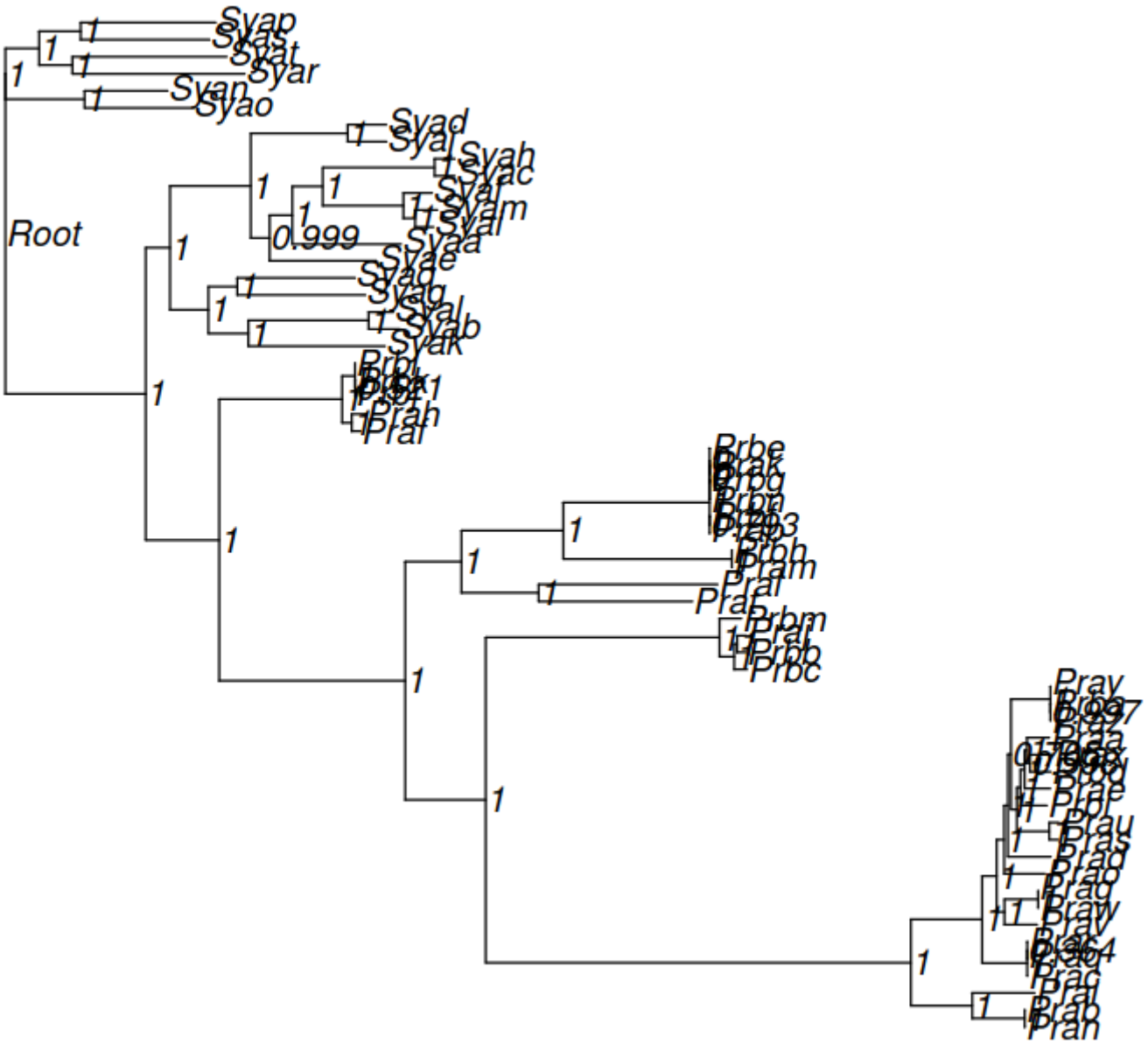


Figure 9 — Arbre d'espèces enraciné obtenu avec **OrthoFinder** à partir des orthogroupes de *Prochlorococcus*. La topologie met en évidence la cohérence des relations évolutives déduites des gènes orthologues, distinguant nettement les clades de haute lumière (HL) et de basse lumière (LL) au sein du collectif.

Fichier localisable depuis :
~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsa/Species_Tree/SpeciesTree_rerooted.pdf

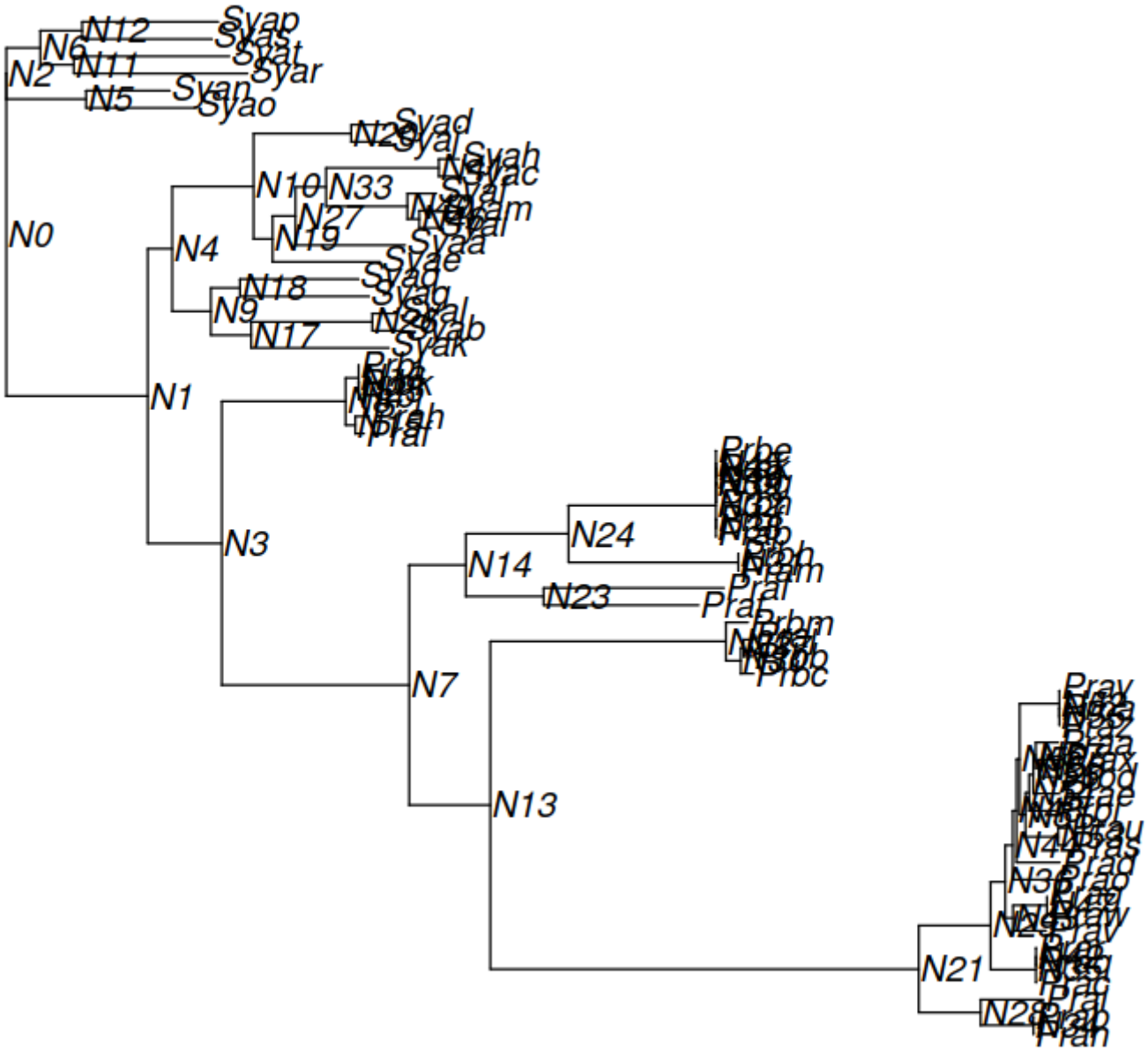


Figure 10 — Arbre d'espèces enraciné de *Prochlorococcus* généré par **OrthoFinder**, avec affichage des étiquettes de nœuds. Cette représentation met en évidence la structure hiérarchique des clades et permet d'associer chaque bifurcation à un ensemble précis d'orthogroupes, facilitant l'interprétation des relations évolutives entre écotypes.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsaroot/Species_Tree/SpeciesTree_rooted_node_labels.pdf

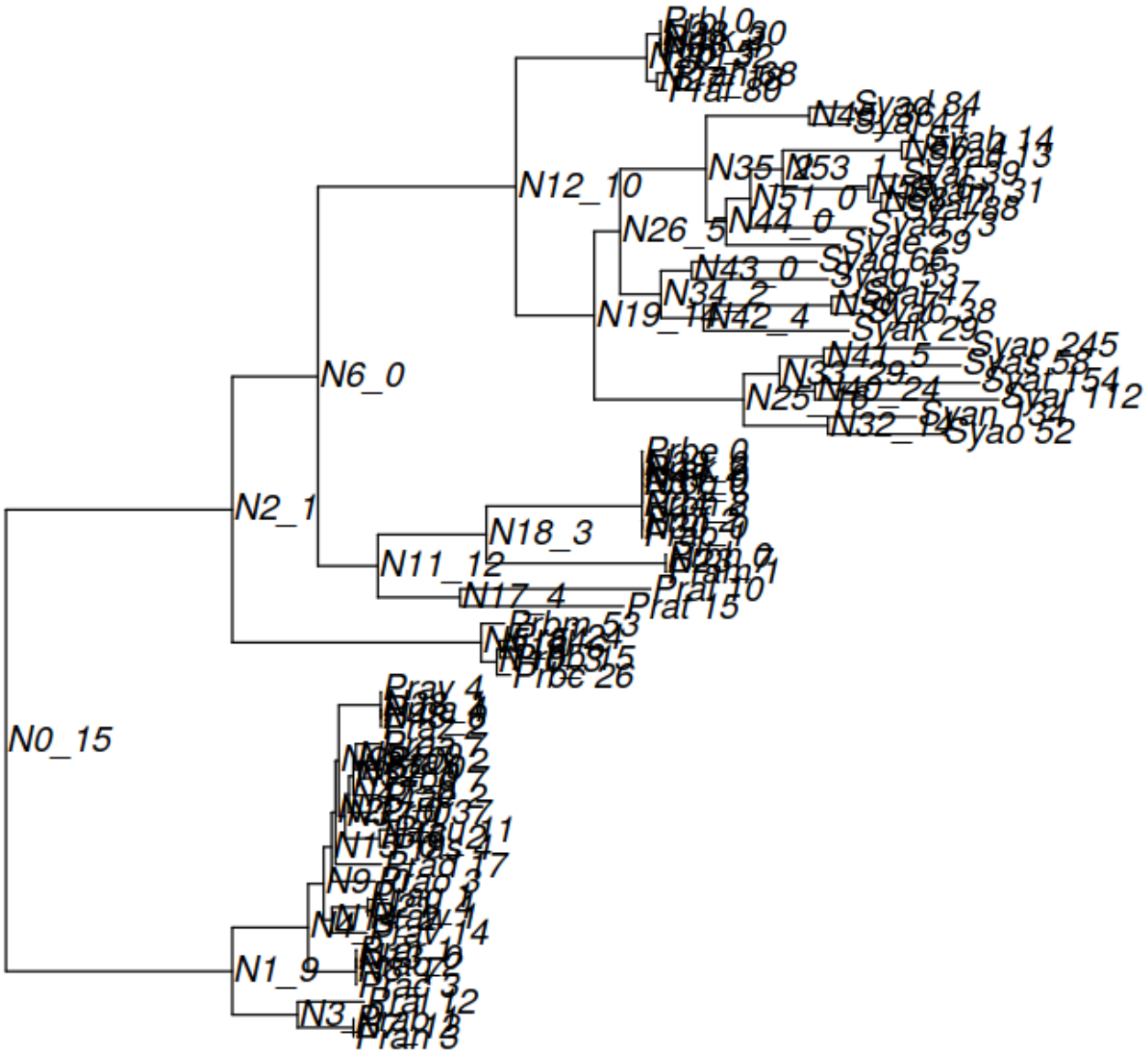


Figure 11 — Arbre d'espèces de *Prochlorococcus* annoté avec les événements de duplication génique identifiés par **OrthoFinder** (seuil de support ≥ 0,5). La figure met en évidence les branches présentant les taux de duplication les plus élevés, suggérant des épisodes d'expansion génétique associés à l'adaptation écologique du collectif.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsa/Gene_Duplication_Events/SpeciesTree_Gene_Duplications_0.5_Support.pdf

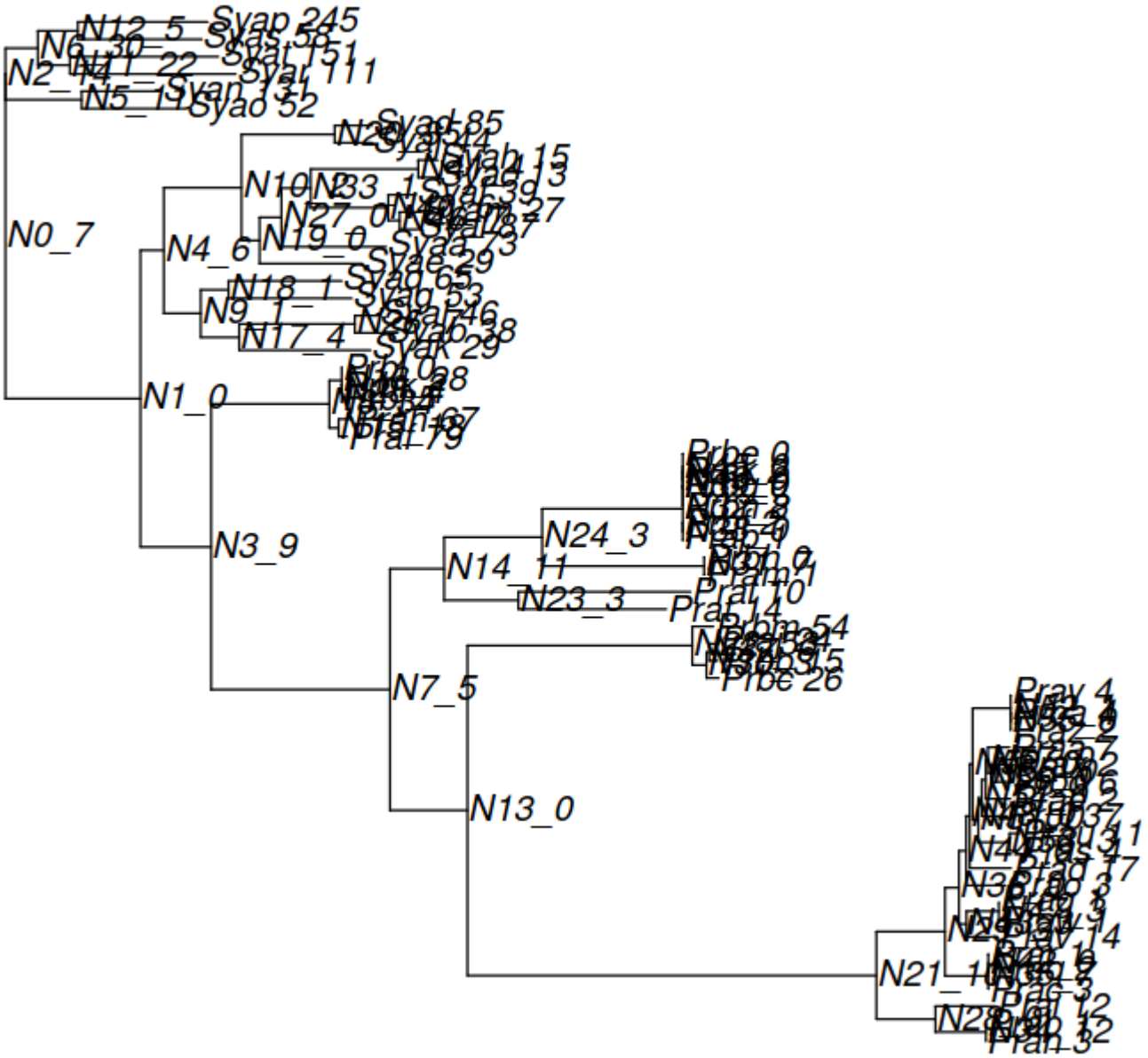


Figure 12 — Arbre d'espèces de *Prochlorococcus* annoté avec les événements de duplication génique détectés par **OrthoFinder** (support ≥ 0,5). Les branches colorées indiquent les lignées ayant connu des duplications significatives, suggérant des épisodes d'expansion fonctionnelle liés à l'adaptation aux différentes niches lumineuses et écologiques.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsaroot/Gene_Duplication_Events/SpeciesTree_Gene_Duplications_0.5_Support.pdf

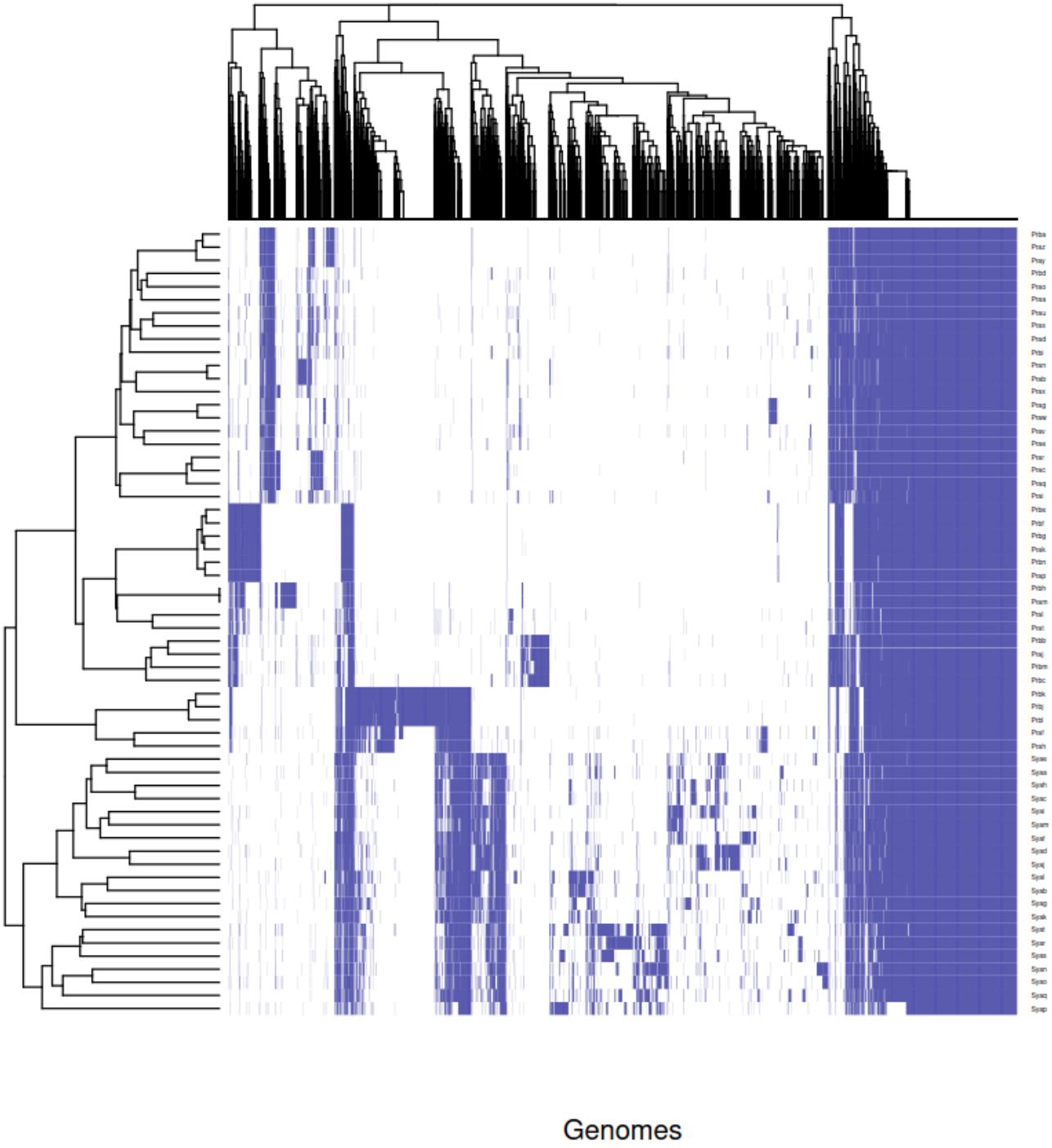


Figure 13 — Matrice de présence et d'absence des orthogroupes identifiés par **OrthoFinder** pour les souches de *Prochlorococcus*. La densité des zones colorées illustre la répartition du génome cœur et du génome accessoire, témoignant d'une forte variabilité génétique entre les différents écotypes du collectif.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/WorkingDirectory/OrthoFinder/Results_ProSynmsa/Orthogroups_matrice01.pdf

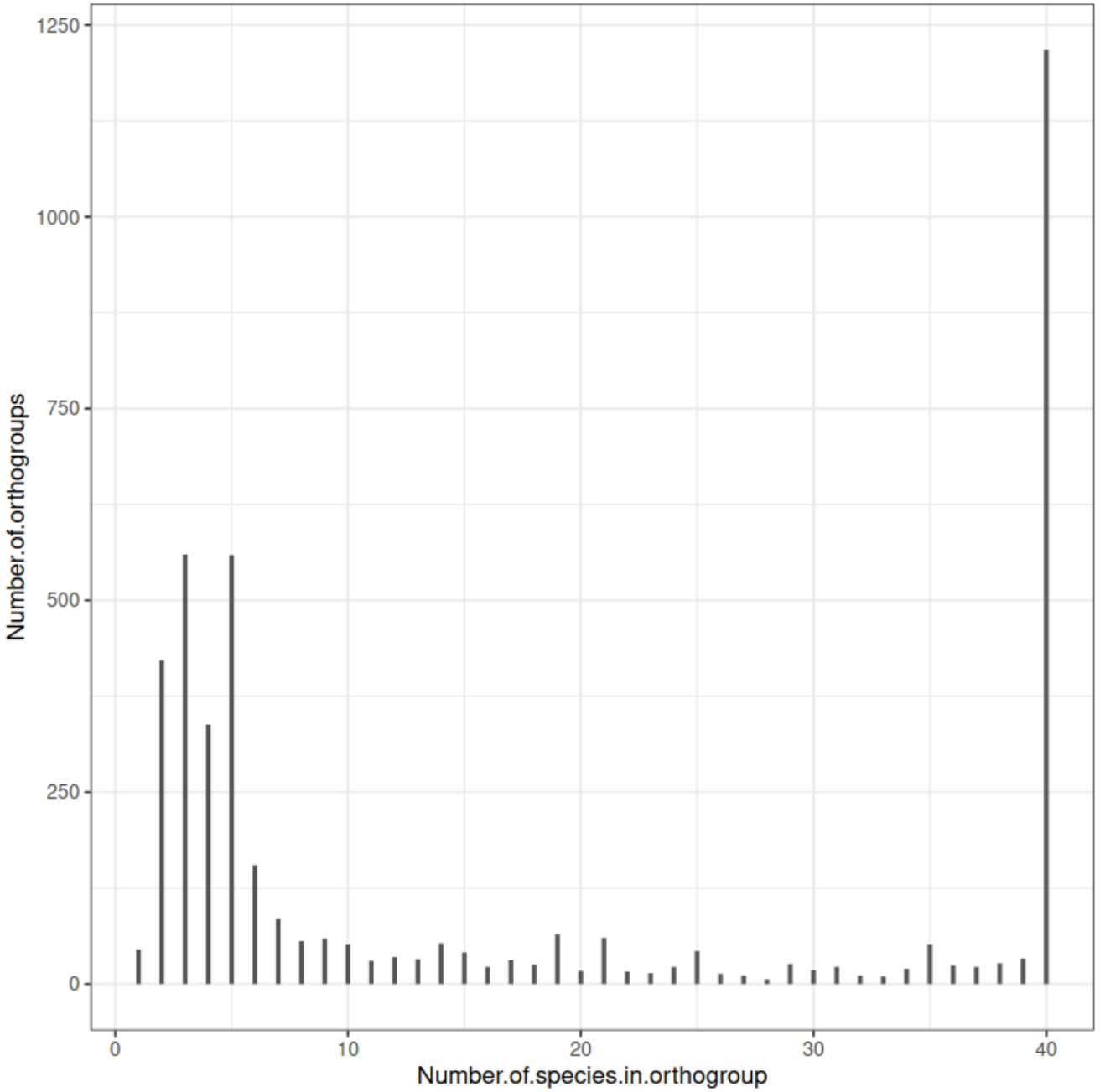


Figure 14 — Histogramme de la distribution du nombre d'espèces par orthogroupe chez *Prochlorococcus*, calculé à partir des résultats d'**OrthoFinder**. La figure illustre la proportion relative de gènes partagés entre les génomes, distinguant le génome cœur (présent dans toutes les souches) du génome accessoire et des gènes spécifiques, caractéristiques d'un pan-génome ouvert.

Fichier localisable depuis : ~/work/OrthoFinder/Prochlorococcus/OrthoFinder/Results_Pro/Comparative_Genomics_Statistics/species_in_OG.pdf

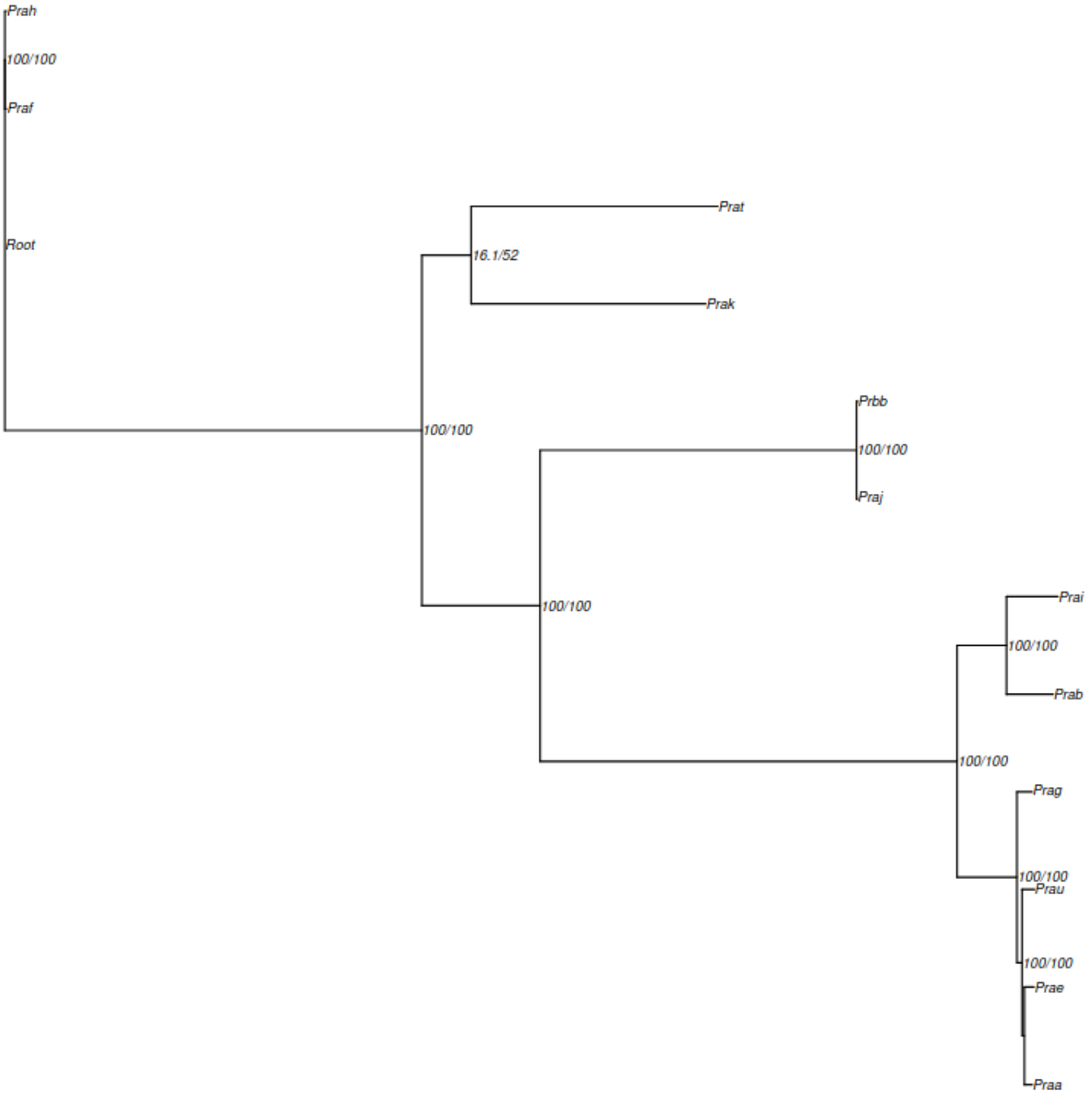


Figure 15 — Arbre phylogénétique des souches de *Prochlorococcus* obtenu par maximum de vraisemblance sous le modèle MFP (ModelFinder Plus) à partir du super-alignement de 31 gènes du génome cœur. L'arbre est enraciné sur les souches profondes *Prah* et *Praf*. Les valeurs de bootstrap indiquent une forte robustesse des principaux clades correspondant aux écotypes HLI, HLII et LL.

Fichier localisable depuis : ~/work/Kettler/phyloG/alignments_reroot_MFP.pdf

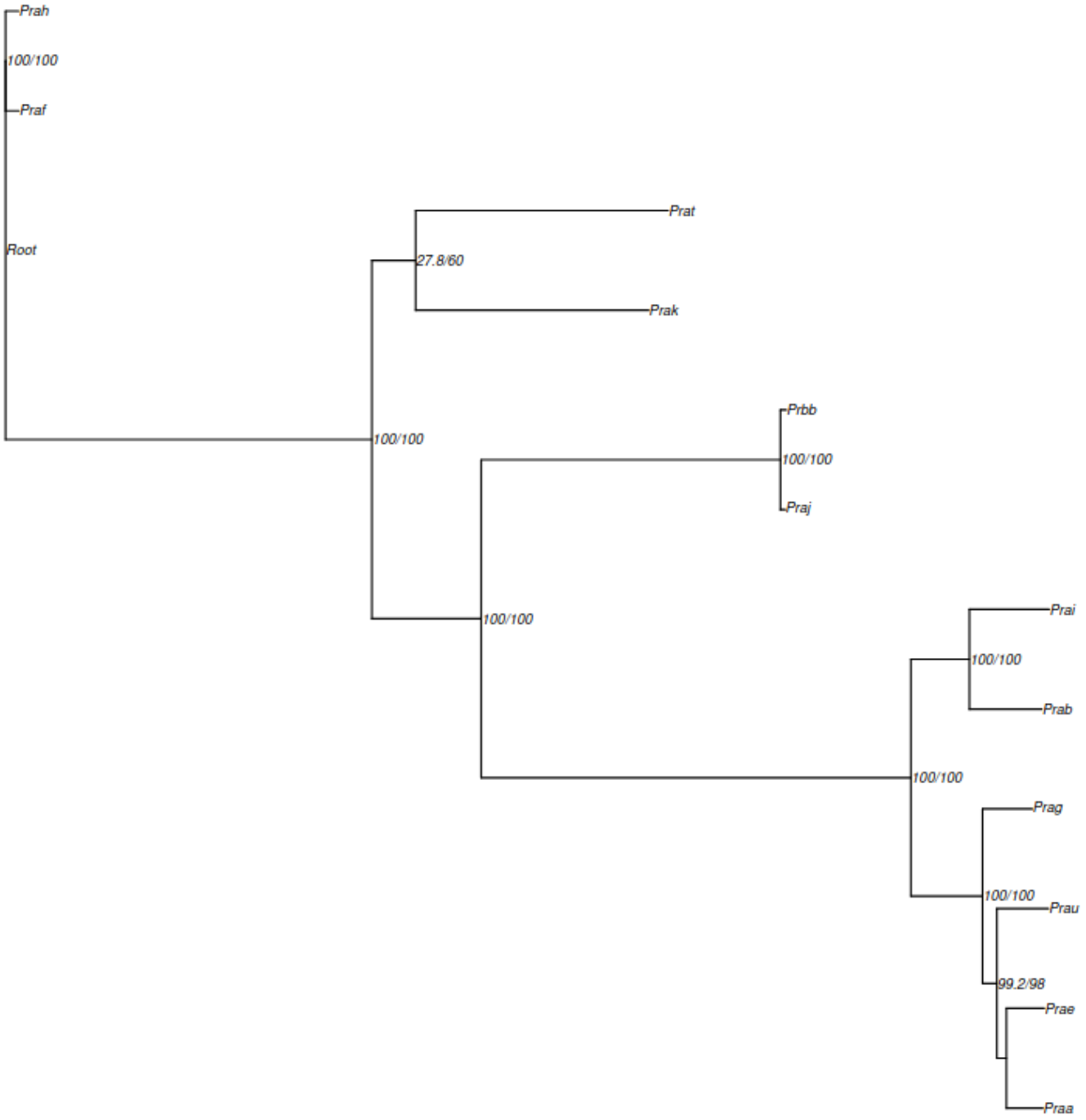


Figure 16 — Arbre phylogénétique des souches de *Prochlorococcus* obtenu à partir du super-alignement protéique sous **IQ-TREE**, utilisant les modèles mixtes **WAG**, **LG**, **JTT**, **LG4M** et **LG4X**. L'arbre reraciné sur *Prah* et *Praf* met en évidence la séparation attendue entre les écotypes de haute lumière (HL) et de basse lumière (LL).

Fichier localisable depuis : ~/work/Kettler/phyloG/alignments_reroot_Prot.pdf

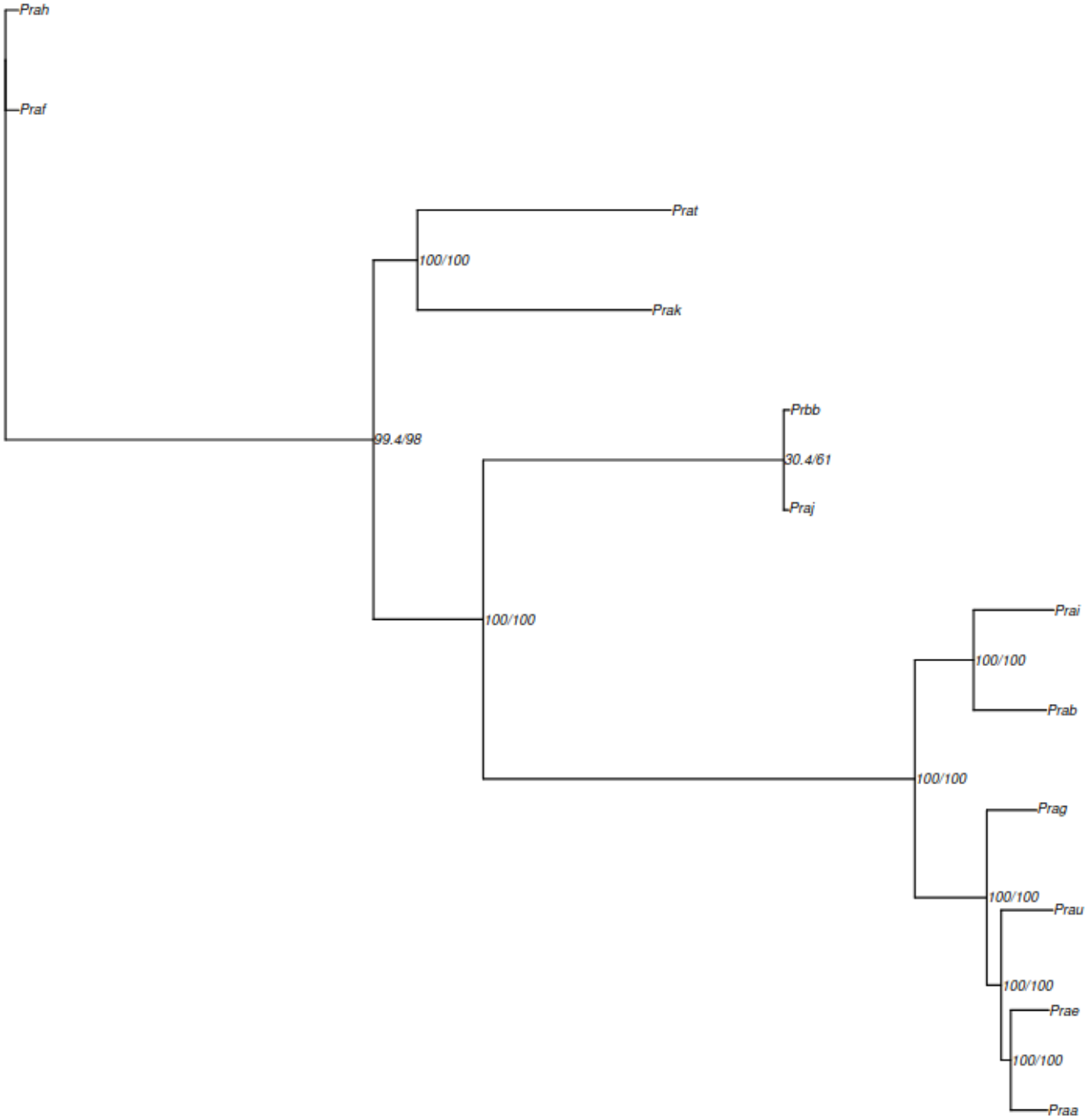


Figure 17 — Arbre phylogénétique des souches de *Prochlorococcus* obtenu à partir du super-alignement protéique, reconstruit par maximum de vraisemblance avec **IQ-TREE** et reraciné à l'aide de **TreeTools**. L'arbre met en évidence une séparation nette entre les clades de haute lumière (HL) et de basse lumière (LL), conforme à la structure écologique connue du collectif *Prochlorococcus*.

Fichier localisable depuis : ~/work/Kettler/phyloG/PEPalignments_reroot_treetools.pdf

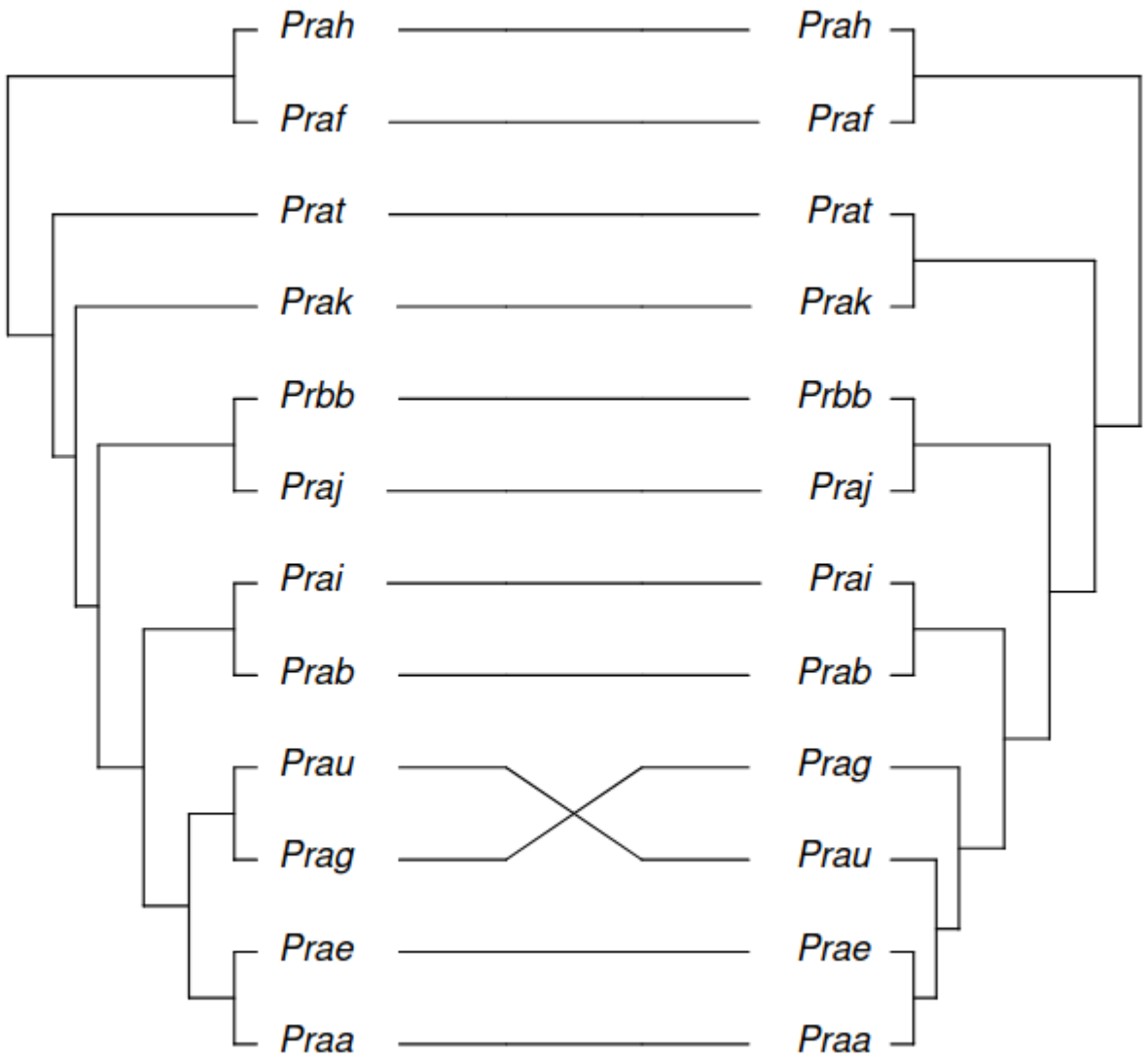


Figure 18 — Représentation comparative (**cophyloplot**) des arbres obtenus à partir des alignements codoniques et protéiques. Les correspondances entre les feuilles montrent une forte cohérence topologique entre les deux approches, confirmant la stabilité des relations phylogénétiques entre les différents écotypes de *Prochlorococcus*.

Fichier localisable depuis : ~/work/Kettler/phyloG/CODON_PEP_cophyloplot.pdf

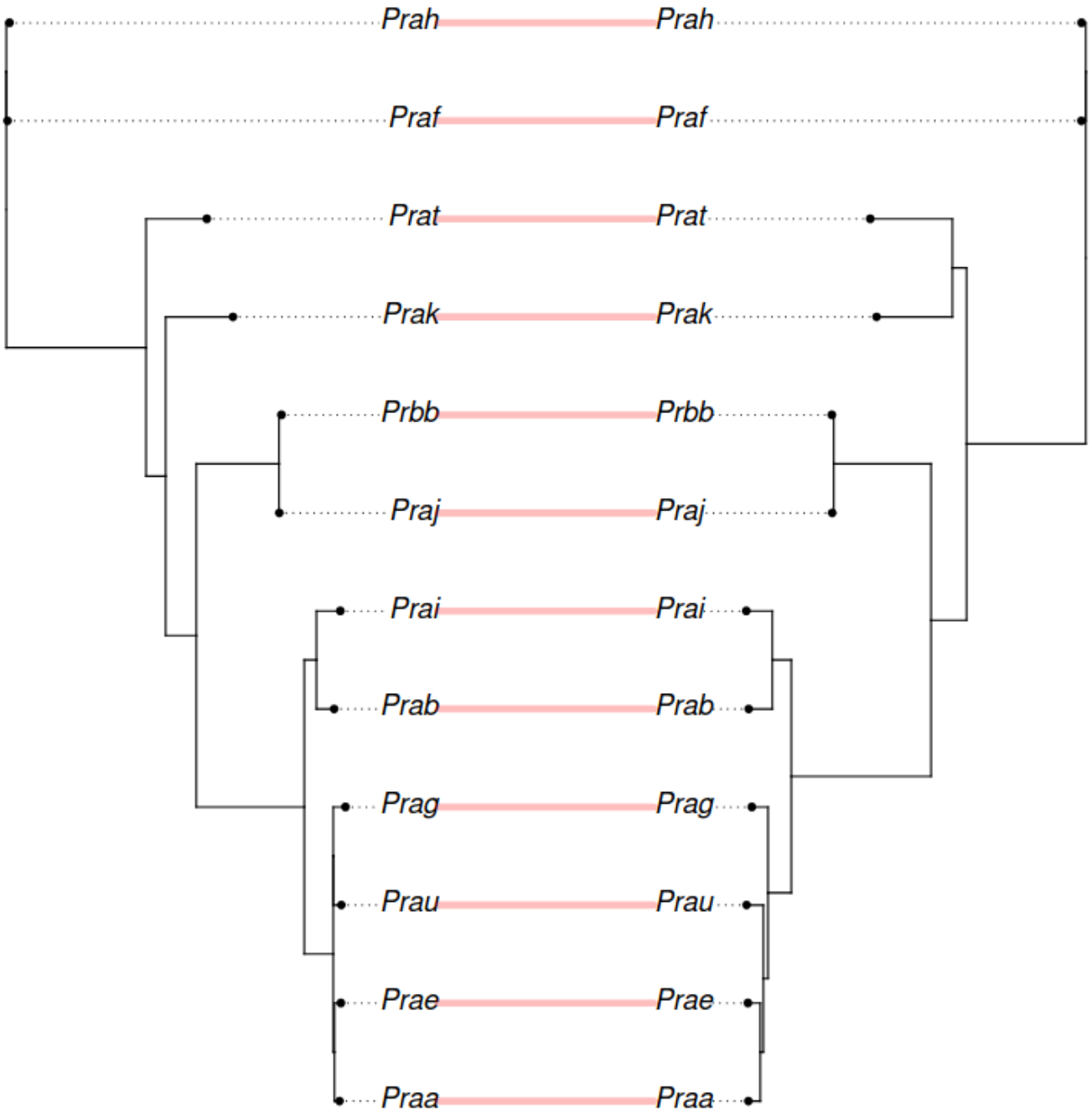


Figure 19 — Visualisation comparative (**cophylogénie**) des arbres obtenus à partir des alignements codoniques et protéiques des souches de *Prochlorococcus*. Les connexions entre les deux arbres illustrent une forte cohérence des relations évolutives, confirmant la stabilité des principaux clades identifiés.

Fichier localisable depuis : ~/work/Kettler/RNAr/ssu_cophylo.pdf

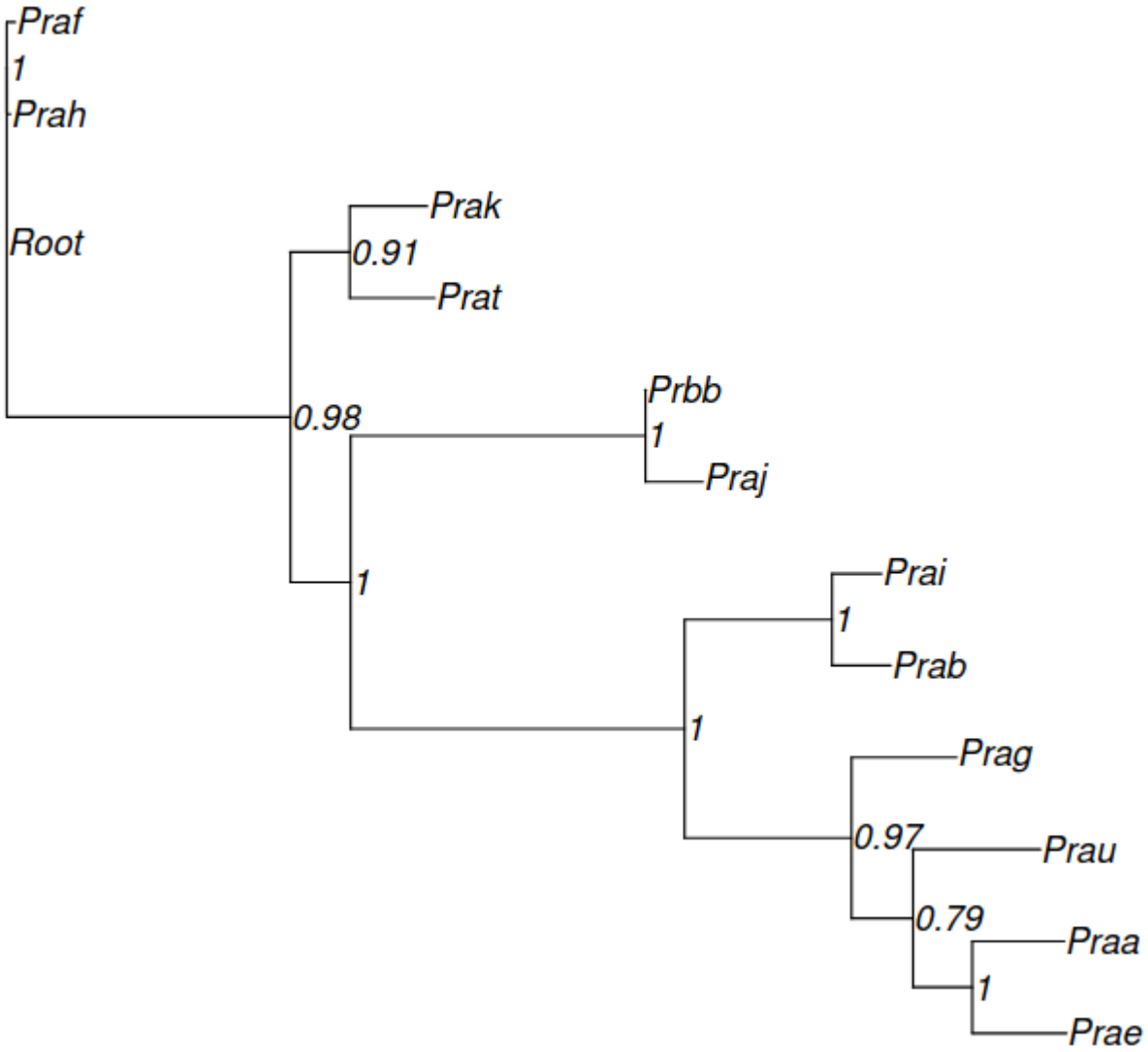


Figure 20 — Super-arbre MRP (Matrix Representation Parsimony) des souches de *Prochlorococcus* obtenu à partir des arbres individuels de gènes protéiques. L'arbre consensus met en évidence la cohérence globale des relations évolutives entre écotypes, avec une distinction marquée entre les clades de haute et basse lumière.

Fichier localisable depuis : ~/work/Kettler/phyloG/MRP_supertrees.pdf

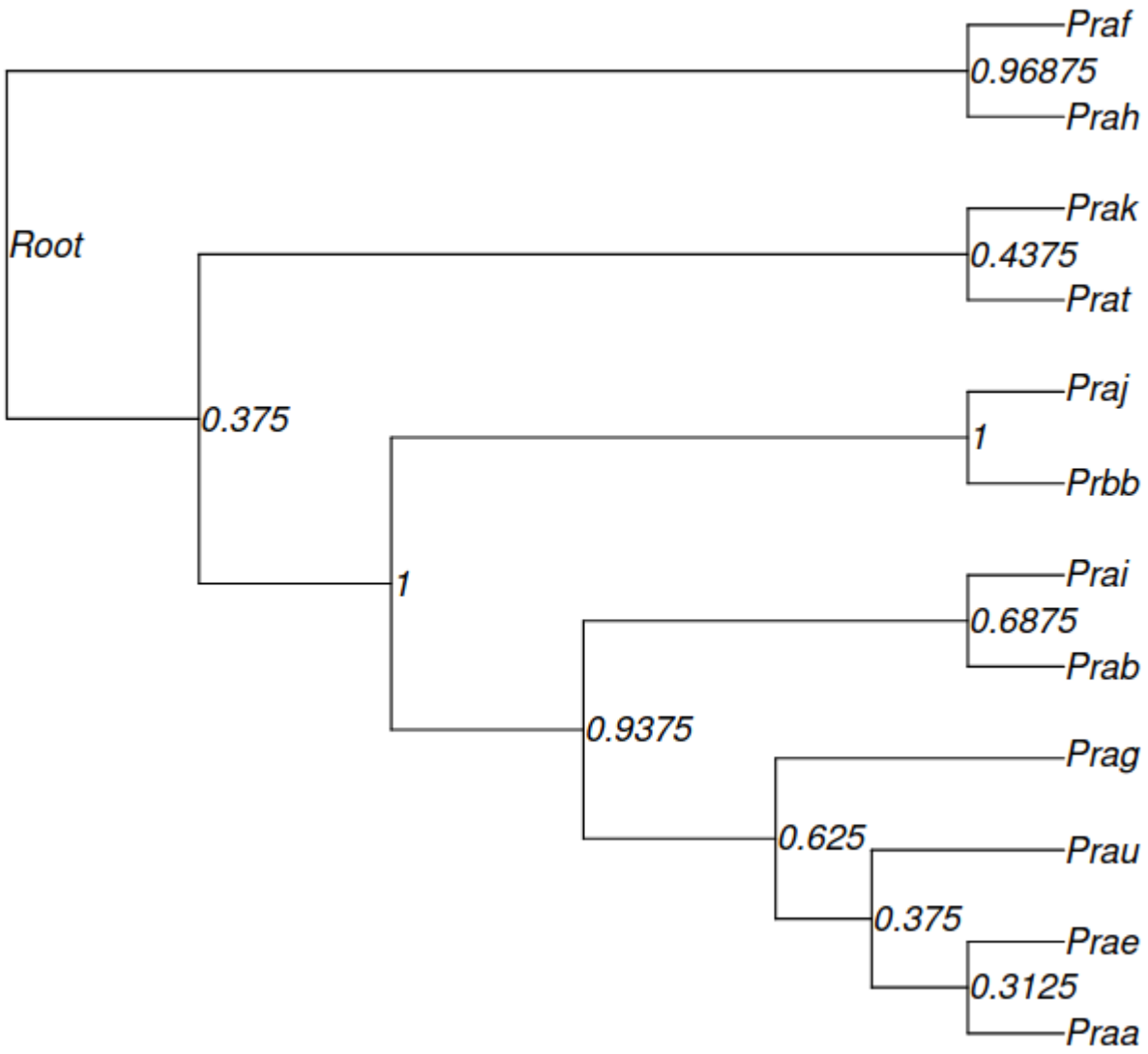


Figure 21 — Arbre consensus des souches de *Prochlorococcus* obtenu à partir des arbres de gènes individuels. La topologie majoritaire reflète les regroupements les plus fréquemment observés, confirmant la cohérence des relations évolutives entre les clades de haute et basse lumière.

Fichier localisable depuis : ~/work/Kettler/phyloG/consus_trees_3.pdf

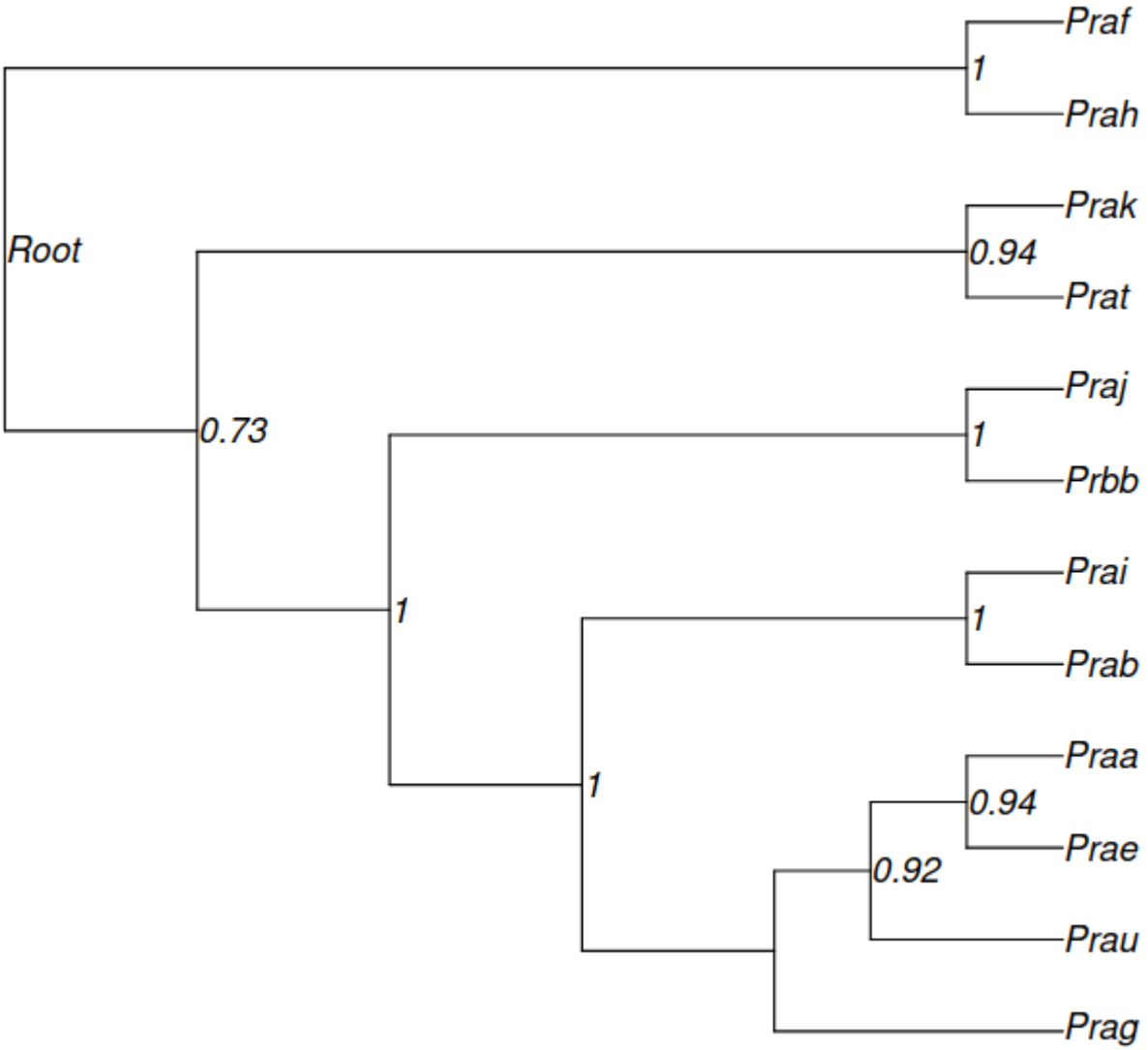


Figure 22 — Arbre d'espèces enraciné obtenu avec **ASTRAL** après suppression des branches présentant un support bootstrap inférieur à 10 %. Cette version épurée met en évidence les relations les plus robustes entre les souches de *Prochlorococcus*, tout en confirmant la séparation entre les écotypes de haute et basse lumière.

Fichier localisable depuis : ~/work/Kettler/phyloG/astral_rooted_supp10percentSupport.pdf

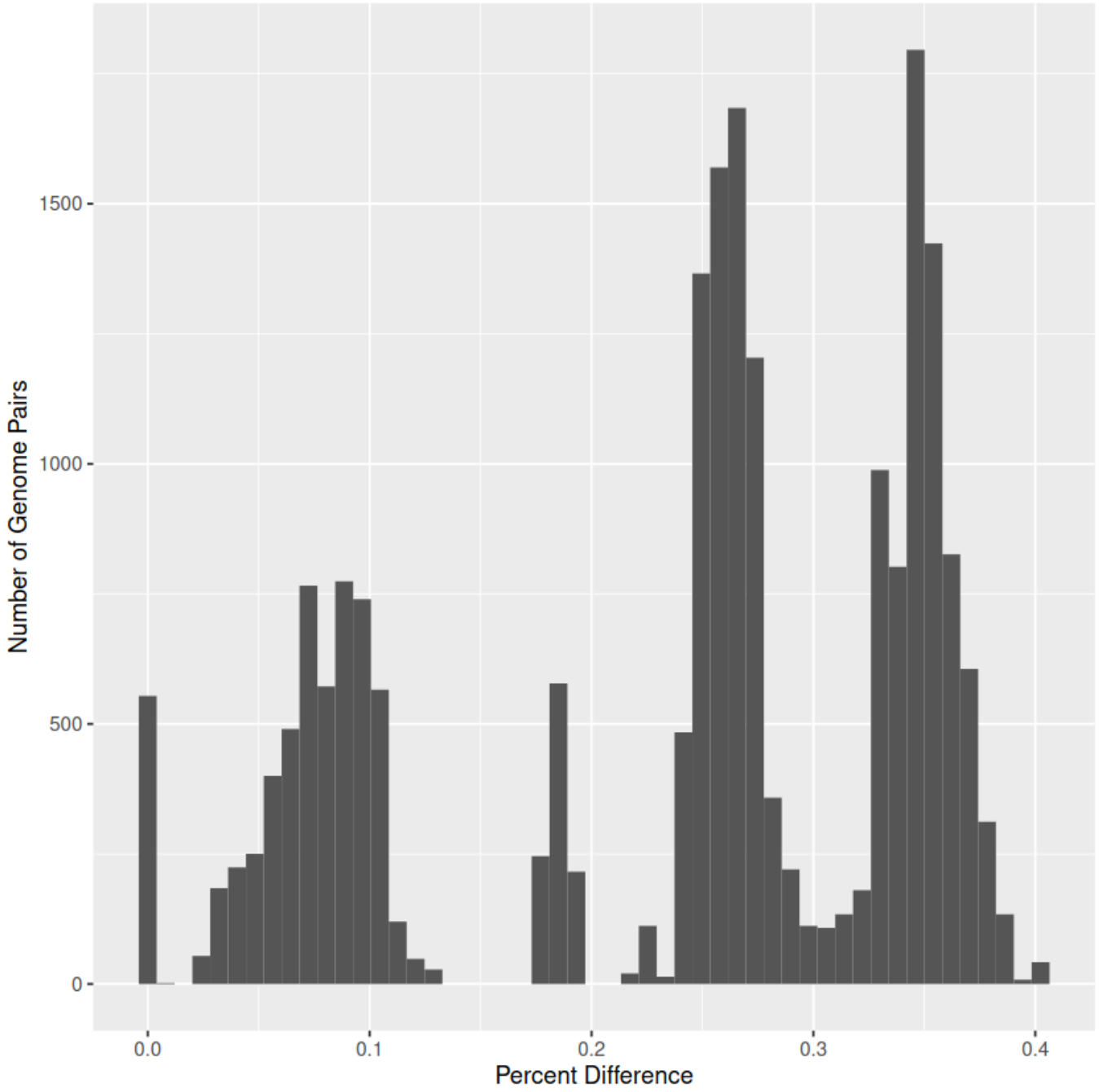


Figure 23 — Histogramme de la distribution de la taille des génomes de *Prochlorococcus* issus de la base **RefSeq**. La figure illustre la variabilité génomique au sein du collectif, caractérisée par une gamme de tailles reflétant l'adaptation des différents écotypes aux environnements marins de haute et basse lumière.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/Prochlorococcus_grrasp.hist.pdf

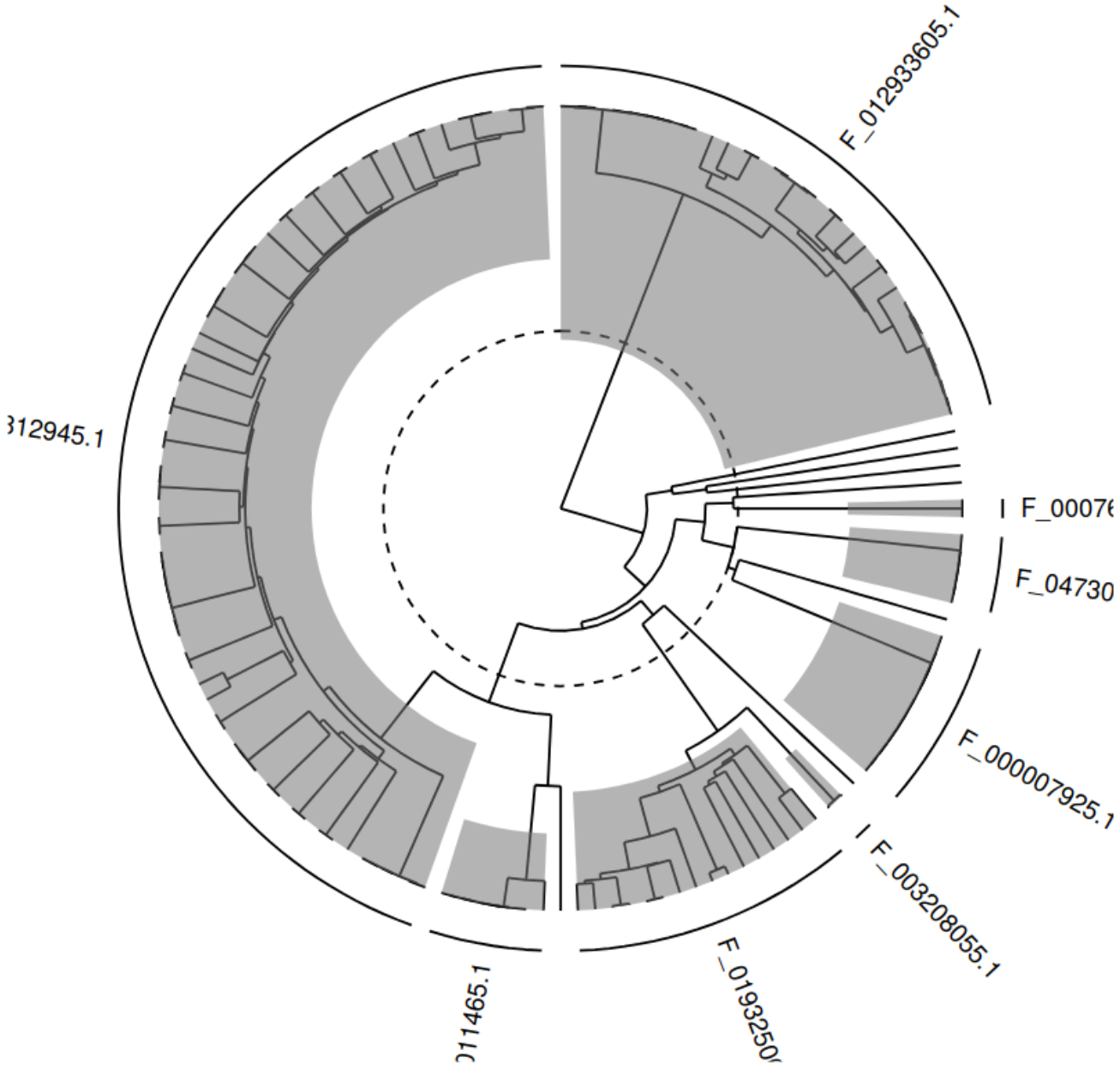


Figure 24 — Arbre de similarité génomique des souches de *Prochlorococcus* obtenu à partir des distances **Mash**. La figure révèle la structuration du collectif en plusieurs clades distincts, correspondant aux grands écotypes connus (HLI, HLII, LL), et illustre la proximité génétique variable entre souches selon leur niche écologique.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/Prochlorococcus_grrasp.clusttree.pdf

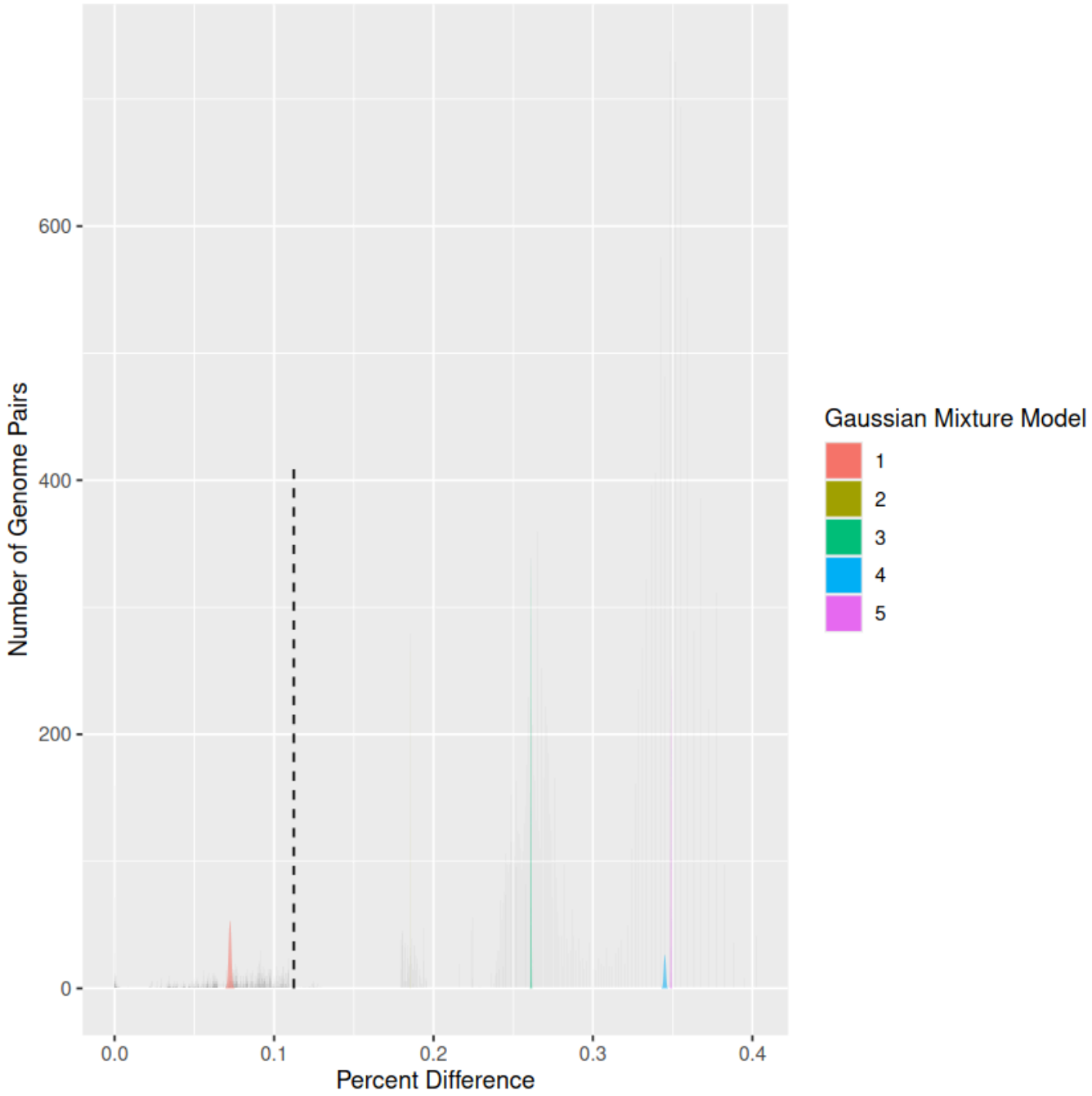


Figure 25 — Classification des génomes de *Prochlorococcus* par **modélisation de mélanges gaussiens (GMM)** à partir des distances Mash. La figure met en évidence plusieurs regroupements correspondant aux grands écotypes du collectif, illustrant la structuration phylogénétique et la différenciation génomique entre les clades de haute et de basse lumière.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/Prochlorococcus_grrasp.gmm.pdf

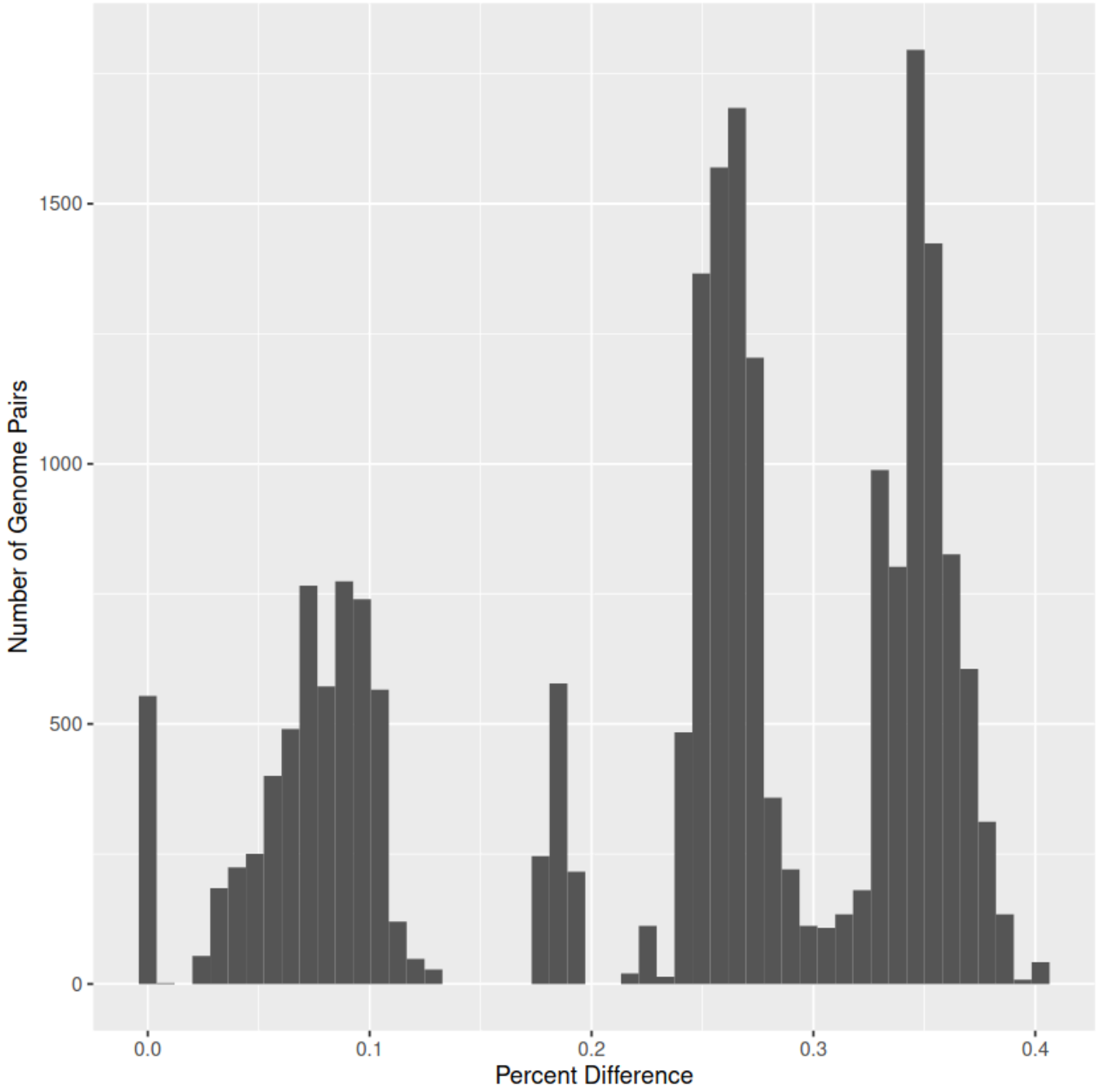


Figure 26 — Histogramme de la complétude et de la contamination des génomes de *Prochlorococcus* évaluées avec **CheckM**. La figure montre que la majorité des génomes présentent une complétude élevée et une contamination faible, garantissant la qualité et la fiabilité du jeu de données utilisé pour les analyses phylogénomiques.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/CheckM_output_folder/Prochlorococcus_grraspQ.hist.pdf

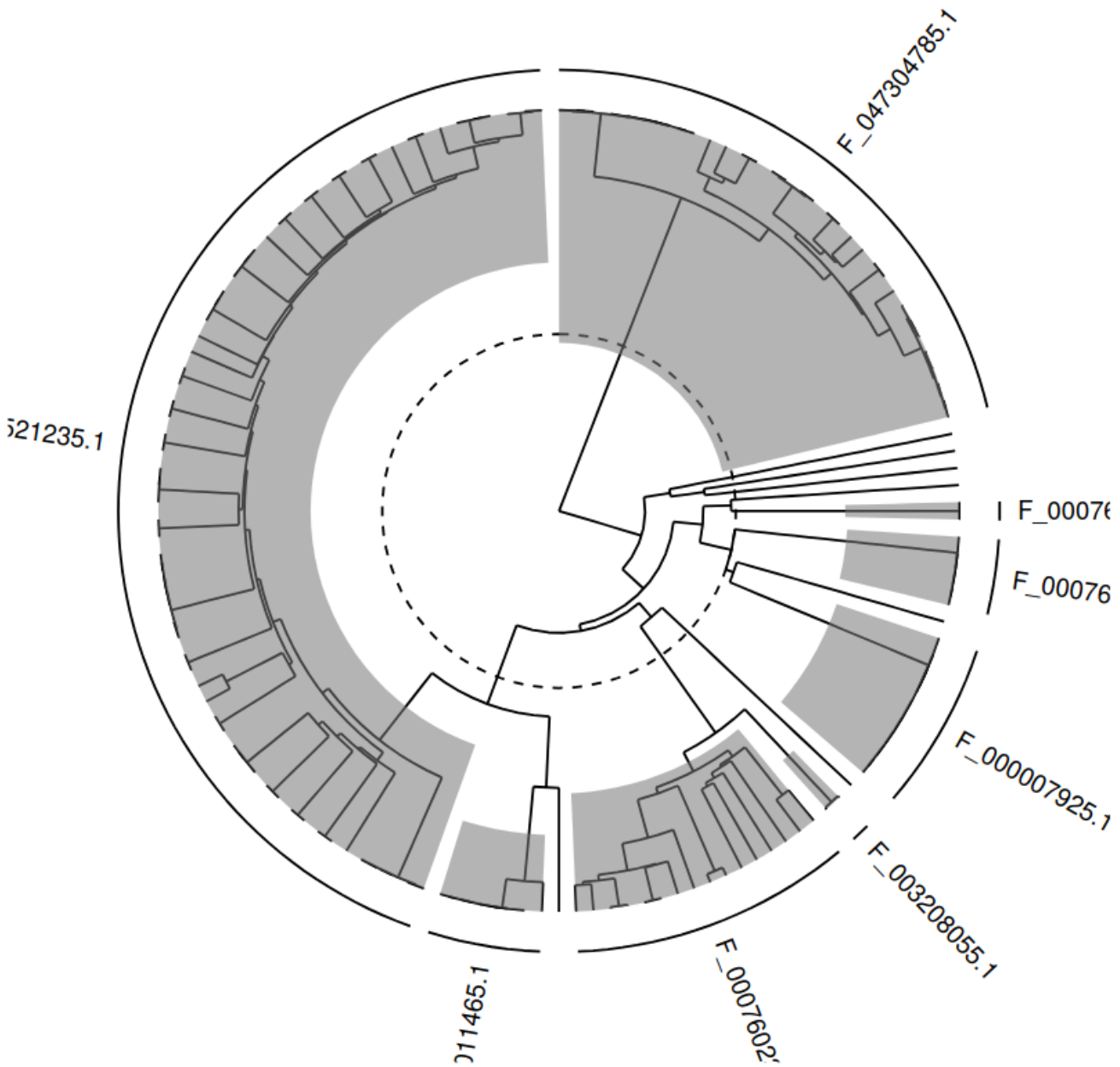


Figure 27 — Arbre de similarité des génomes de *Prochlorococcus* basé sur les scores de complétude et de contamination issus de **CheckM**. La figure met en évidence des regroupements cohérents entre génomes de qualité comparable, permettant d'identifier les assemblages fiables à conserver pour les analyses évolutives.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/CheckM_output_folder/Prochlorococcus_grraspQ.clusttree.pdf

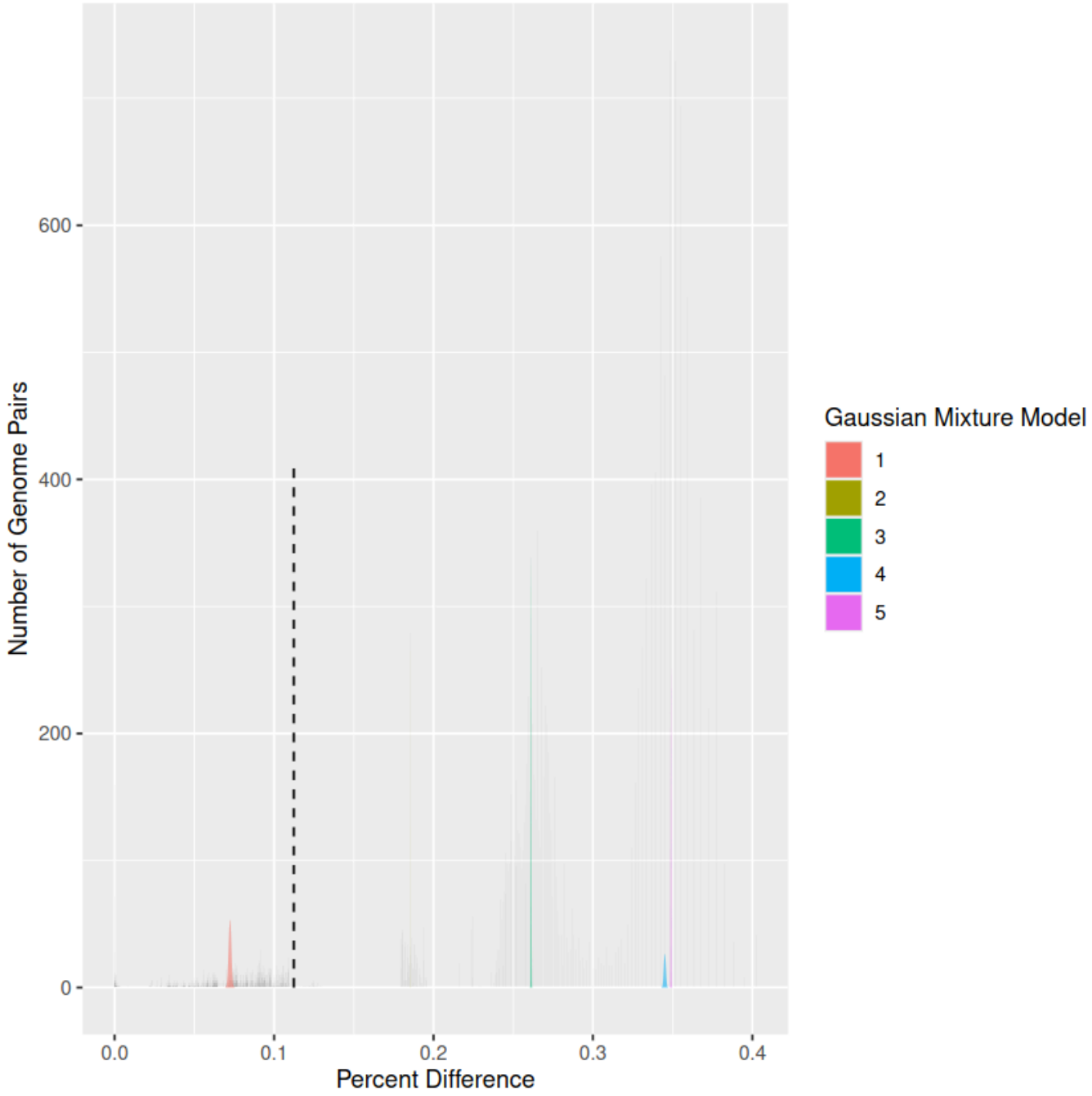


Figure 28 — Regroupement des génomes de *Prochlorococcus* selon leurs scores de complétude et de contamination évalués par **CheckM**, modélisé par un **mélange gaussien (GMM)**. La figure distingue clairement les génomes de haute qualité des assemblages partiels ou contaminés, assurant une sélection rigoureuse pour les analyses comparatives ultérieures.

Fichier localisable depuis : ~/work/Prochlorococcus/RefSeq/CheckM_output_folder/Prochlorococcus_grraspQ.gmm.pdf

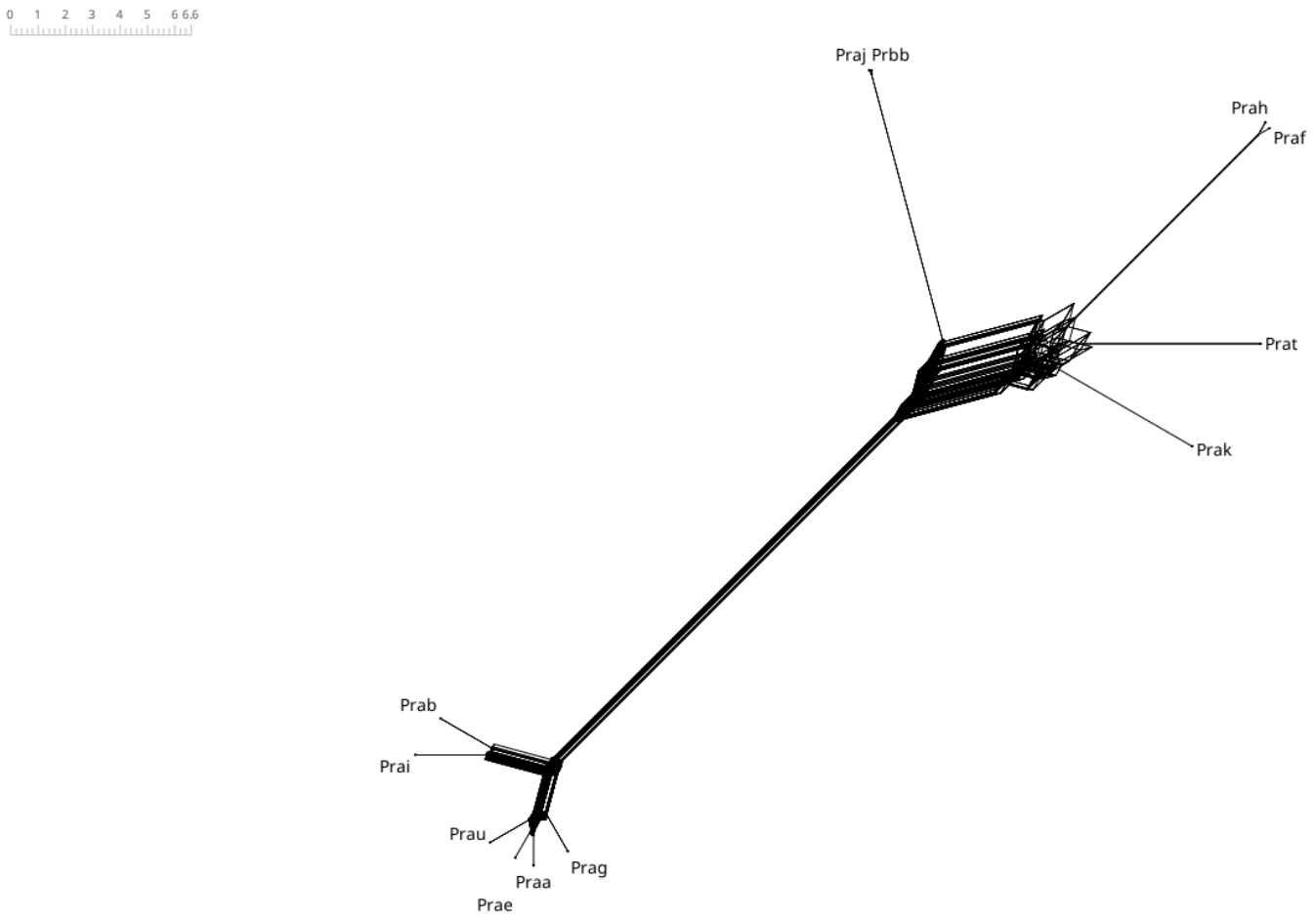


Figure 29 — Réseau phylogénétique obtenu avec SplitsTree à partir du fichier *alltrees.tree.nex.* La figure visualise les relations évolutives entre les souches de Prochlorococcus sous forme de réseau de compatibilité, mettant en évidence d'éventuels signaux conflictuels entre gènes et la complexité des relations dues aux duplications ou transferts horizontaux.

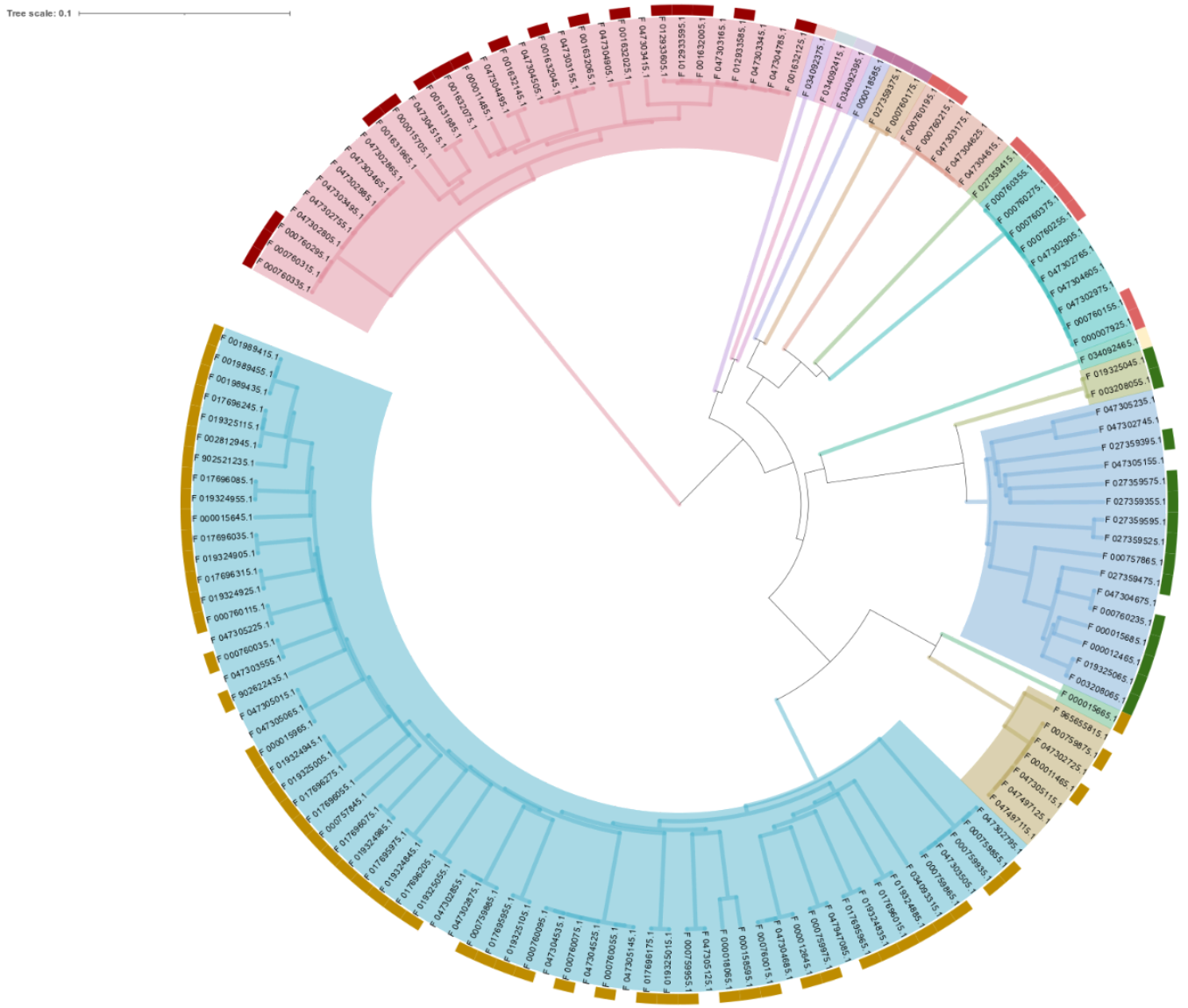


Figure 30 — Arbre phylogénétique des génomes de *Prochlorococcus* obtenu à partir des distances Mash (`Prochlorococcus_grrasp.tree.txt`) et visualisé sous **iTOL** à l'aide des fichiers d'annotation `Prochlorococcus_grrasp.itol.txt` et `bac120_taxonomy_ACsuffix_group_genera_iTOL.txt`. La topologie circulaire représente la diversité génétique du collectif *Prochlorococcus*, tandis que la **couronne extérieure** indique les genres définis par la taxonomie génomique récente.
Les couleurs distinctes associées à ces groupes soulignent leur **divergence phylogénétique** et leurs **adaptations écologiques spécifiques**, notamment aux gradients de lumière et de nutriments dans l'océan. Cette figure illustre ainsi la redéfinition taxonomique du collectif historique *Prochlorococcus marinus* en plusieurs genres génétiquement cohérents.