Study ot the coevolution between antimicrobial peptides and peptide transporters in legume-rhizobium symbiosis

Amira Boukherissa

THÈSE DE DOCTORAT

NNT : 2024UPASB043

université PARIS-SACLAY

Study of the coevolution between antimicrobial peptides and peptide transporters in legume-rhizobium symbiosis

Étude de la coévolution entre peptides antimicrobiens et transporteurs de peptides dans le cadre la symbiose rhizobium-légumineuses

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°567, Sciences du végétal : du gène à l'écosystème (SEVE)
Spécialité de doctorat : Évolution
Graduate School : BioSpheRA. Référent : Faculté des Sciences d'Orsay

Thèse préparée dans les unités de recherche **Institute for Integrative Biology of the Cell I2BC (Université Paris-Saclay, CEA, CNRS)** et **Ecologie, Systématique et Évolution (Université Paris-Saclay, CNRS, AgroParisTech)** sous la direction de **Benoît ALUNNI**, directeur de recherche et le co-encadrement de **Ricardo RODRIGUEZ DE LA VEGA**, ingénieur de recherche.

Thèse soutenue à Paris-Saclay, le 30 septembre 2024, par

Amira BOUKHERISSA

Composition du Jury

Membres du jury avec voix délibérative

KARINE ALIX Professeur, AgroParisTech, GQE Le Moulon, Gif-sur-Yvette, Université de Paris-Saclay	Présidente
Delphine CAPELA Directrice de recherche CNRS, LIPME, Castanet-Tolosan, Université de Toulouse	Rapportrice & Examinatrice
Andrei N. LUPAS Professeur, Max Planck Institute for Biology Tübingen	Rapporteur & Examinateur
Jesús MONTIEL GONZÁLEZ Maître de conférence, National Autonomous University of Mexico, Mexico	Examinateur

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences du végétal : du gène à l'écosystème (SEVE)

Titre : Étude de la coévolution entre peptides antimicrobiens et transporteurs de peptides dans le cadre la symbiose rhizobium-légumineuses

Mots clés : symbiose, phylogénie moléculaire, évolution, transporteur ABC, peptides antimicrobiens

Résumé :

Les légumineuses présentant une carence en azote peuvent entrer en interaction symbiotique avec des bactéries du sol fixatrices de N2 appelées rhizobia. Dans cinq clades de légumineuses, une stratégie d'exploitation appelée différenciation terminale des bactéroïdes (TBD) a évolué dans laquelle les rhizobiums subissent une différenciation extrême. Les bactéries terminalement différenciées sont plus grandes, polyploïdes, ont une membrane perméabilisée, et sont meilleures à la fixation de N2, fournissant un retour sur investissement plus élevé pour la plante. Nous savons que dans deux clades, IRLC (par exemple, Medicago spp.) et Dalbergioïds (par exemple, Aeschynomene spp.), ce processus de différenciation est déclenché par un ensemble de peptides antimicrobiens végétaux apparemment non apparentés avec une activité antimicrobienne à la membrane connue sous le nom de peptides Nodule-spécifiques Cystéine-Riche (NCR). À son tour, les rhizobia exposés au stress provoqué par les NCRs nécessitent un transporteur de peptides ABC de la famille BacA pour faire face à ce stress. Cependant, si des peptides NCR ou des peptides similaires sont également trouvés dans d'autres clades où la TBD se produit et la relation évolutive entre ces peptides reste inconnue. Dans ce projet, nous avons testé l'hypothèse d'une coévolution convergente entre les différents clades de légumineuses et leur rhizobia engagés dans ce programme de différenciation, tant au niveau phénotypique que moléculaire. Pour ce faire, nous avons combiné des analyses d'évolution moléculaire avec des tests fonctionnels, fournissant ainsi des connaissances expérimentales sur la question fondamentale de la contingence et de répétabilité en évolution tout en générant simultanément de nouveaux outils pour concevoir une symbiose plus efficace.

Title: Study of the coevolution between antimicrobial peptides and peptide transporters in legume-rhizobium symbiosis

Keywords: molecular phylogeny, symbiosis, ABC transporter, evolution, antimicrobial peptides

Abstract:

Legume plants under nitrogen deficiency can enter a symbiotic interaction with N_2-fixing soil bacteria called rhizobia. In five legume clades, an exploitive strategy called Terminal Bacteroid Differentiation (TBD) has evolved in which rhizobia undergo extreme differentiation. Terminally differentiated bacteria are larger, polyploid, have a permeabilized membrane, and are better at N_2 fixation, providing a higher return on investment for the plant. We know that in several members of the distantly related Inverted Repeat Lacking Clade (IRLC, e.g., *Medicago* spp.) and the Dalbergioid clade (e.g., *Aeschynomene* spp.), this differentiation process is triggered by a set of apparently unrelated plant antimicrobial peptides with membrane damaging activity known as Nodule-specific Cysteine-Rich (NCR) peptides. In turn, rhizobia exposed to NCR stress requires an ABC peptide transporter of the BacA family to cope with this stress. However, whether NCR peptides or similar peptides are also found in other clades where this occurs and the evolutionary relation among these peptides remain unknown. In this project, we tested whether NCR peptides and BacA peptide transporters evolved independently in the different legume clades that induce TBD and their rhizobia, implying convergent coevolution, both at phenotypic and molecular levels. We combined molecular evolution analyses with functional assays, thus providing experimentally informed knowledge on the fundamental question of the part of contingency and repeatability in evolution while simultaneously generating new tools to engineer a more efficient symbiosis.

FOR MY FATHER

"You were always my greatest teacher and supporter. You taught me to be curious and work hard to achieve my goals. You gave me all the confidence, support, and love to be a good person and realize my dreams. I am very grateful for everything you have done for me. I am sure that if you were here, you would have been proud of your daughter."

Acknowledgments

I want to acknowledge all the people who have contributed to my thesis work in different ways. This project was only successful with their participation.

First, I would like to thank Dr. Benoît Alunni, my thesis director, for his valuable help on plant biology and writing parts of my project. I learned a lot from his expertise in plant biology and his experience in plant-microbe interactions. Thanks to him, I gained independence because he always believed in me and my skills.

Dr. Ricardo Rodriguez de la Vega, my supervisor, thanks a lot for your invaluable help with the bioinformatics and molecular evolution part of my project. I would like to thank him for his valuable guidance and assistance in using the relevant methods to analyze the complex datasets I generated. I also express my sincere gratitude for this presence and his constructive feedback. Although he helped me to extract meaningful conclusions from my results, he always pushed me to find the answers, solve the issues, and interpret my results by myself.

I also sincerely thank Dr. Jacqui Shykoff, my second supervisor, for her important help on the evolutionary part of my project, which helped me to understand the evolutionary context of my results. She also contributed to the writing part of the project, where she helped me edit my research papers, and she always pushed me to write my results to improve the clarity of my work. In addition to her contribution to the important scientific parts of my project, I would also like to thank her for her important support during the challenging times of my thesis.

Many thanks to our collaborators from the diCenzo lab at Queen's University in Canada. Their contribution was invaluable and extended over different parts of the project. They provided me with important bioinformatic pipelines, additional data, and bacterial strains that enriched my analysis, which led us to collaborate and work together. I am expressing my specific gratitude to Dr. George DiCenzo for the opportunity he gave me to work for three months in his lab. Thanks a lot for your help in getting the Mitacs funding, for your help with administrative stuff in Canada, for your supervision, and for always being present to answer my questions and give me clear answers during my thesis journey. Since he was the most present person in my thesis project, his contribution was invaluable, and I learned a lot from his side. I would also like to thank the diCenzo lab students, especially Nick, Mia, and Rui, for their help and collaboration. I learned a lot from their expertise and enjoyed working with them.

I would like to extend my gratitude to my colleagues and friends, Sara and Roza, whose invaluable support and guidance have been instrumental throughout my PhD. They were very important during this journey. I am incredibly lucky to have gained not only knowledgeable colleagues but also important friends.

I would like to thank my team leaders, Dr. Peter Mergeart and Dr. Tatiana Giraud, for trusting me and giving me this opportunity to work in their teams using the CNRS 80Prime funding they obtained. I would also like to thank Dr. Peter Mergeart for always being supportive and encouraging—many thanks for trusting in my expertise and including me in your project. I would also like to thank Dr. Emanuele Biondi for his presence and for sharing his invaluable knowledge in microbiology. He was always ready to help and discuss science. He is very generous and friendly.

I would like to thank Dr. Tania Timchenko for her valuable help in the wet biology part of my project. She gave me essential molecular biology protocols and instructions on using them and managing the samples.

I would also like to thank my thesis committee members, Pr. Alessandra Carbone, Pr. Peter Tiffin, Dr. George diCenzo, and Dr. Jeanne Ropars, for their time and effort in reviewing and evaluating my work. I value the constructive critique and insightful recommendations they provided, which have enhanced the quality of my research.

All my deep gratitude to Guislaine Refregier for her support and guidance during the challenging moments of my thesis journey. I am grateful for her wisdom and how she helped me maintain my well-being during these difficult moments.

Many thanks to all the members of my two teams, PBI from I2BC and GEE from ESE, who have been wonderful colleagues and friends. They have helped me with various aspects of the project. A particular acknowledgment goes to Corinne Foucault, who helped me collect the samples. They have also shared their ideas and opinions and provided me with moral and emotional support. I enjoyed working with them and learning from them.

Last, I would like to thank my family, my mother, Lydia, Madjda, Yasmine, Romaissa, Mira, and Imen, who have been my pillars of strength and motivation. Their constant presence and support made my life more meaningful. I dedicate this thesis to the two persons who are not present now but who marked my life: my grandmother and my father.

Résumé en Français

1. Introduction

L'évolution est-elle reproductible ? est une question cruciale en biologie évolutive. La question de savoir si l'adaptation est contingente ou reproductible est d'une importance capitale pour notre compréhension de la nature de l'évolution (D. Collins, 1990). Les études sur les populations naturelles ont montré que l'adaptation à des niches écologiques similaires a conduit à l'évolution convergence des caractères phénotypiques, mais peu d'études ont examiné si la convergence s'est également produite aux niveaux moléculaire et génétique, alors que de telles études fonctionnelles pourraient permettre de répondre à la question suivante : les convergences phénotypiques découlent-elles de voies évolutives similaires et utilisent-elles des mécanismes moléculaires similaires ?

La symbiose légumineuse-rhizobium est un excellent exemple pour étudier la coévolution convergente, montrant comment différents couples d'espèces (hôte-symbionte) peuvent parvenir indépendamment à des solutions similaires grâce à l'évolution convergente.

Dans des conditions de déficit en azote, les légumineuses (*Fabaceae*) peuvent entrer en symbiose avec des bactéries du sol fixatrices d'azote (N_2) connues sous le nom de rhizobia, qui font partie des protéobactéries (alpha et bêta). Au cours de cette interaction, la légumineuse forme des nodosités racinaires où les rhizobia sont hébergés de façon intracellulaire sous forme de structures appelées bactéroïdes, qui fixent l'azote atmosphérique et transfèrent l'ammoniac à la plante. À leur tour, les plantes fournissent à ces micro-symbiontes une source de carbone et un apport nutritif. Cette interaction s'initie après la reconnaissance mutuelle de la plante hôte avec un partenaire bactérien compatible par un échange de molécules de signalisation entre les deux partenaires (Oldroyd, 2013). Premièrement, les flavonoïdes sont sécrétés par la légumineuse. Une fois internalisés dans la bactérie, ils induisent la production de lipooligosaccharides appelés facteurs Nod. La perception de ces facteurs par la légumineuse induit le processus de nodulation, initié par l'attachement des bactéries aux poils racinaires, qui se recourbent enfermant une micro-colonie qui progresse au sein de cordons d'infection (IT pour "Infection Thread"). En parallèle, des divisions cellulaires au niveau du péricycle et du cortex interne permettent la mise en place d'un méristème qui donnera naissance à la nodosité. Les cordons d'infection se ramifient au sein du primordium nodulaire acheminant ainsi les rhizobia jusqu'à leur relargage intracellulaire dans les cellules végétales (Gage, 2004). A

l'intérieur de la nodosité et plus précisément à l'intérieur du compartiment subcellulaire appelé symbiosome, les rhizobiums deviennent des bactéroïdes fixateurs d'azote. Dans certains programmes symbiotiques, les bactéroïdes restent similaires aux bactéries en culture, leur forme, leur métabolisme et leur physiologie ne sont pas modifiées pendant la symbiose (Lamouche, Bonadé-Bottino, et al., 2019; Oono & Denison, 2010). Cependant, chez certaines légumineuses comme *Medicago truncatula* et *Aeschynomene evenia* appartenant aux clades des IRLC et Dalbergioïdes, respectivement, les bactéroïdes ont une morphologie modifiée où ils deviennent allongés ou sphériques, respectivement (Lamouche, Bonadé-Bottino, et al., 2019). Cet état irréversible des bactéroïdes augmente l'efficacité symbiotique et est atteint par un processus appelé différenciation terminale des bactéroïdes (TBD) (Haag & Mergaert, 2020; Lamouche, Bonadé-Bottino, et al., 2019; Oono & Denison, 2010). Les bactéroïdes différenciés sont plus allongés, polyploïdes et ont des membranes perméabilisées (Alunni & Gourion, 2016; Mergaert et al., 2006). Ce processus de différenciation est induit par de petits peptides végétaux appelés NCR (*Nodule-specific Cysteine-Rich*) qui sont fortement exprimés dans les nodules des légumineuses qui induisent cette différenciation (Pan & Wang, 2017; Van de Velde et al., 2010). Ces peptides sont composés d'un peptide signal qui permet leur sécrétion et d'un peptide mature composé de 20-50 acides aminés dont 4 ou 6 cystéines conservées qui forment deux ou trois ponts disulfures (Maróti et al., 2015). Les séquences des acides aminés des peptides matures sont très variables, à l'exception des cystéines conservées (Mergaert et al., 2003). Le nombre de peptides NCR chez les légumineuses varie de 7 (*Glycyrrhiza uralensis*) à 700 (*Medicago truncatula*) (Montiel et al., 2017; Young et al., 2011). Selon leur point isoélectrique, les peptides NCR peuvent être classés comme cationiques, neutres et anioniques. Les NCR cationiques présentent une activité antimicrobienne responsable de la perméabilisation de la membrane des bactéroïdes (Maróti et al., 2011, 2015). En outre, il a été suggéré que les peptides NCR provoquent un changement de cycle cellulaire en symbiose, où il a été démontré que NCR247 peut inhiber la division cellulaire bactérienne en interagissant avec la protéine FtsZ (Farkas et al., 2014). Les bactéries rhizobia peuvent supporter le stress provoqué par les peptides NCR et prévenir les dommages de la membrane à l'aide de transporteurs ABC appelés BacA ou BacA-like qui sont essentiels pour établir une symbiose efficace uniquement avec les légumineuses qui induisent une différenciation terminale des bactéroïdes (Glazebrook et al., 1993; Guefrachi et al., 2015; Haag et al., 2011). BacA est un transporteur de peptides, dont le mutant de délétion est incapable de transporter des peptides NCR et montre une sensibilité élevée aux peptides NCR cationiques provoquant la mort rapide des rhizobia et l'incapacité de fixer l'azote. Bien que BacA et BclA transportent principalement les peptides NCR dans les

rhizobia, des résultats récents suggèrent que le transporteur YejABEF peut également contribuer à l'importation de peptides NCR en montrant que les mutants yejAEF sont sensibles à au moins un peptide NCR *in vitro* (Nicoud et al., 2021).

Nous savons que ce processus de différenciation terminale des bactéroïdes se produit dans cinq clades de légumineuses (Génistoïdes, Mirbéloïdes, IRLC, Milletioïdes et Dalbergioïdes) (Oono et al., 2010). Dans deux de ces clades, IRLC et Dalbergioïdes, ce processus de différenciation dépend des peptides NCR sécrétés par les légumineuses et des transporteurs BacA dans les rhizobia. Cependant, la présence de peptides NCR dans d'autres clades où la TBD a été identifiée ou n'a pas encore été recherchée, leur identité moléculaire, leur évolution, et l'implication de BacA et BclA dans la différenciation et leur évolution restent inconnues. Ainsi, dans ce projet, nous avons cherché à déchiffrer l'histoire évolutive des peptides NCR et des transporteurs BacA et essayé de savoir si une coévolution convergente conduit l'évolution des peptides NCR et des transporteurs BacA.

Donc, pour répondre à cette question principale de mon projet de thèse, nous avons combiné des analyses bioinformatiques d'évolution moléculaire et des expériences fonctionnelles en laboratoire pour répondre aux questions suivantes :

- La différenciation terminale des bactéroïdes dans les clades sous-étudiées est-elle également due à la production de peptides NCR ?
- Les peptides NCR sont-ils recrutés à partir des mêmes familles de gènes dans les clades IRLC, Dalbergioïdes et autres ?
- Les protéines bactériennes BacA ont-elles suivi la même voie évolutive pour la résistance aux NCR dans les rhizobia ? D'où ont-ils évolué ? Quelle est leur fonction ancestrale ?
- Les peptides antimicrobiens végétaux et les systèmes de résistance bactérienne co-évoluent-ils ?

2. Résultats

A. La convergence et la divergence des peptides antimicrobiens riche en cystéines spécifiques aux nodosités

Pour savoir s'il y a d'autres peptides NCR dans les clades IRLC et Dalbergioides, s'il y a des peptides NCR dans les espèces de légumineuses non étudiés qui induisent la TBD, s'il n'y a

pas de peptides NCR chez les espèces de légumineuses qui n'induisent pas la TBD, étudier leur identité moléculaire et inférer leur histoire évolutive, nous avons effectué une comparaison inter- et intra-clade de ces peptides antimicrobiens.

Nous avons commencé par la comparaison intra-clade des peptides NCR connus dans les génomes disponibles de légumineuses IRLC (*Medicago truncatula, Medicago sativa, Cicer arietinum* et *Pisum sativum*) et Dalbergioides (*Arachis hypogaea* et *Aeschynomene evenia*). En utilisant des méthodes d'homologie, d'orthologie et de clusterisation, nous avons regroupé dans des clusters, ces peptides NCR bien connus des clades IRLC et Dalbergioides. Brièvement, nous avons recueilli les dernières versions des données génomiques et protéomiques des espèces de légumineuses mentionnées ci-dessus avec des peptides NCR connus. Ensuite, nous avons effectué une analyse de similarité à l'aide de la commande blastp (Camacho et al., 2009) pour reporter la similarité entre toutes les séquences de toutes les espèces et nous avons utilisé ces scores pour définir des ensemble d'orthologues à l'aide du logiciel orthAgogue (Ekseth et al., 2014). Les orthologues ont été regroupés en clusters en utilisant l'algorithme de Clustering de Markov (MCL) (Van Dongen, 2008). Tous les groupes orthologues contenant des NCR ont été clade-spécifique. Parmi les 1523 peptides NCR dans le clade IRLC, 1492 ont été attribués à 651 groupes d'orthologues. Parmi eux, 203 (568 NCRs) étaient des clusters contenant exclusivement des peptides (NCR-exclusive) NCR, et les 448 (924 NCRs) autres étaient des groupes d'orthologues mixtes avec au moins un NCR et un non-NCR (NCR-mixte). Dans les Dalbergioides, 117 peptides NCR sur 155 ont été classifiés en 20 clusters. Parmi ces 20 clusters, 7 (40 NCRs) étaient des clusters NCR-exclusive et 13 (77 NCRs) NCR-mixte. Pour les analyses qui suivent, nous avons considéré seulement les clusters avec au moins deux peptides NCR avec un peptide signal, où nous avons exclu les clusters NCR-monotypique avec uniquement un seul NCR dans le cluster. Par conséquent, nous avons gardé 385 clusters de NCRs chez les IRLC et 11 clusters NCR chez les Dalbergioïds. De ces clusters, nous avons extrait les séquences d'ADN et d'acides aminés en utilisant des scripts python et nous avons gardé que les séquences qui ont un peptide signal prédit par le logiciel signalP (Petersen et al., 2011). Ensuite, nous avons construit des profils HMM (Hidden Markov Model) à partir des alignements de séquences d'acide aminés (codon-based) de ces clusters. Nous avons utilisé SPADA (P. Zhou et al., 2013) (Small Peptide Alignment Discovery Application) pour identifier les peptides NCR dans toutes les espèces de légumineuses où les données génomiques et transcriptomiques des nodosités sont disponibles (21 espèces appartenant à 6 différents clades). SPADA est un outil de recherche de gènes basé sur l'homologie avec une puissance

spécifiquement améliorée dans la détection et l'identification de gènes avec un ou deux exons et avec un peptide signal. Nous avons exécuté SPADA trois fois séparément pour chaque génome, une fois en utilisant les profils NCR des IRLC, une fois en utilisant les profils NCR des Dalbergioides et une fois en utilisant les profils CRP, qui sont des peptides végétaux riches en cystéines que nous avons utilisés parce que nous n'étions pas certains que nos profils clade-spécifiques puissent capturer tous les NCRs dans d'autres clades. Nous avons ensuite filtré les peptides prédits en fonction de leur longueur, de leur composition en cystéines et de leur expression dans les nodosités. Nous avons ensuite effectué une étape supplémentaire pour les peptides NCR trouvés avec des profils CRP où nous les avons recherchés par rapport à nos profils IRLC et Dalbergioides et les avons classifiés dans le cluster dont le profil possède le meilleur score.

Ensuite, pour étendre notre ensemble de données avec un autre clade de légumineuses non étudiée qui induit la TBD, nous avons généré un ensemble de données RNA-seq de nodules et racines de la légumineuse *Indigofera argentea* appartenant aux *Indigofereae*. Nous avons effectué un assemblage *de novo* de transcriptomes de racines et de nodosités à partir des données RNA-seq (Illumina) que nous avons générées pour *I. argentea* et de données RNA-seq brutes disponibles publiquement pour deux espèces de *Lupinus* du clade Génistoïde (*L. luteus* et *L. mariae-josephae*). Afin de confirmer que *I. argentea* induit la TBD, nous avons quantifié la quantité d'ADN et la taille des bactéroïdes dans les nodosités d'*I. argentea* en utilisant la cytométrie en flux. Cette analyse a montré une polyploïdisation des bactéroïdes avec des pics à 3C.

Selon les mesures de qualité utilisées (BUSCO, alignement et statistiques), les assemblages obtenus sont de très bonne qualité. L'identification des peptides NCR dans ces trois espèces a été effectuée avec SPADA, comme expliqué ci-dessus. Cette analyse nous a permis d'identifier de nouveaux peptides NCR chez 14 espèces de légumineuses qui induisent la TBD, y compris des espèces pour lesquelles les répertoires NCR n'ont jamais été décrits (par ex. 12 NCRs chez *I. argentea* et entre 36 et 87 NCRs chez les Génistoïdes), et de nouveaux NCR dans des clades bien étudiés (par ex. 13% à 70% des NCR chez six espèces IRLC sont nouvellement identifiées ici). Alors que presque tous les peptides NCR nouvellement identifiés dans les IRLC et Génistoïdes ont été classifiés avec nos clusters NCR connus, dans les Dalbergioides, seulement 24 à 32% des NCR ont été classés, et seulement un NCR Indigoféroïdes identifié a été classifié avec un cluster Dalbergioïdes. L'analyse transcriptomique d'*I. argentea* a montré que les 12 gènes codants des peptides NCR sont fortement induits dans les nodosités et que l'un d'eux

appartient aux dix transcrits les plus exprimés. De plus, les abondances de transcrits calculées à l'aide de TPM (Transcripts Per Million) entre les trois espèces de *Lupinus* montrent une expression significativement plus faible des peptides NCR dans *L. albus* que les deux autres espèces de *Lupinus*. En effet, *L. albus* a moins de peptides NCR que les deux autres.

Les peptides NCR sont des peptides courts constitués de 20 à 50 acides aminés dans le peptide mature, et les séquences sont très divergentes. Cette diversification rapide des peptides NCR réduisant la similarité entre les séquences a probablement caché l'origine évolutive de ces peptides et rendu l'inférence de leur histoire évolutive en utilisant l'analyse phylogénétique traditionnelle très difficile. Par conséquent, afin de mieux comprendre l'évolution des peptides NCR, nous avons utilisé la classification et la phylogénétique structurelle pour étudier les peptides NCR au niveau de leur structure 3D. Nous avons prédit les structures 3D de 390 peptides NCR classifiés (un par cluster), 27 peptides NCR non-classifiés de Génistoïdes et Indigoféroïdes, et 48 défensines de quatre clades de légumineuses qui induisent la TBD (ayant un score pLDDT > 70), en utilisant Alphfold2 (Jumper et al., 2021). Nous avons utilisé Foldseek (van Kempen et al., 2024) pour regrouper ces structures en 23 superclusters, dont neuf étaient de petits clusters clades-spécifiques, et les 14 autres étaient inter-clades. Par exemple, cette analyse nous a permis de regrouper les clusters cationiques d'IRLC et de Génistoïdes dans le même supercluster (SC156), y compris le NCR343 qui a été identifié comme essentiel pour une symbiose efficace chez *M. truncatula* et les défensines se regroupent dans un supercluster avec quelques Génistoïdes, Dalbergioïdes et l'un des plus abondants peptide NCR chez les Indigoféroïdes. Le peptide NCR247, le mieux caractérisé chez *M. truncatula,* appartient à un supercluster monotypique et atypique composé de seulement 5 séquences, exclusivement trouvées dans les espèces *Medicago.* La comparaison entre les scores de TM (Template Modelling) et les identités de séquence à l'intérieur du plus grand supercluster a montré que les structures 3D des peptides NCR sont relativement conservées malgré la divergence de leurs séquences, justifiant pourquoi les relations évolutives étaient cachées en utilisant les approches basées sur les séquences. Afin de déchiffrer l'évolution des peptides NCR et des défensines et d'avoir une meilleure vue d'ensemble de nos superclusters, nous avons aussi utilisé Foldtree (Moi et al., 2023), une approche de phylogénie structural qui utilise les distances entre les structures pour construire un arbre phylogénétique. Conformément à la présence de quelques Dalbergioïdes et Indigoféroïdes dans le supercluster des défensines, l'arbre phylogénétique structural a regroupé les superclusters des défensines et des Dalbergioïdes, séparément des autres superclusters regroupant des IRLC-Génistoïdes.

Afin de valider l'approche utilisée pour prédire les peptides NCR et tester leurs fonctions, nous avons synthétisé neuf peptides NCR de différents clades et de différents superclusters. Il a été rapporté récemment que le peptide NCR247 peut se lier à l'hème et le séquestrer, facilitant l'importation de fer par les rhizobia (Sankari et al., 2022). Afin de vérifier si d'autres peptides NCR pourraient également se lier ce cofacteur, nous avons mesuré l'absorption de la lumière selon les longueurs d'onde (nm). Il est intéressant de noter que parmi les neuf NCR testés, un seul peptide NCR de *M. truncatula* se lie bien à l'hème à 420 nm, tandis que le peptide NCR247 de *M. truncatula* se lie à 360 et 450 nm. Notamment, parmi les sept peptides induisant des changements du cycle cellulaire chez *S. meliloti*, le peptide NCR d'*I. argentea* appartenant au supercluster des défensines avec un motif à huit cystéines a induit l'amplification génomique de l'ADN, une autre caractéristique des peptides NCR après avoir été fortement et différentiellement exprimés dans les nodules. Ce résultat valide notre approche de recherche et de classification des peptides NCR et soutient l'hypothèse selon laquelle les peptides NCR ont évolué à partir du répertoire immunitaire des défensines dans le clade des Indigoféroïdes.

Bien que cela ne soit pas entièrement résolu, deux scénarios évolutifs ont émergé de ces résultats, révélant l'histoire évolutive cachée des NCRs. D'une part, les peptides NCR de Dalbergioïdes et d'Indigoféroïdes ont évolué à partir du pool de défensines. D'autre part, compte tenu du fait que les clades IRLC et Génistoïdes sont relativement éloignés, les peptides NCR dans ces deux clades sont recrutés indépendamment par évolution convergente, suivi d'une expansion et une rapide diversification de cette famille au sein des IRLC. En effet, les peptides NCR de ces deux clades se regroupent dans les mêmes clusters et superclusters avec quelques peptides IRLC monotypiques.

B. La distribution taxonomique des transporteurs de peptides antimicrobiens SbmA/BacA et BacA suggère un recrutement indépendant et une évolution convergente dans les interactions hôte-microbe

Comme mentionné précédemment, les transporteurs BacA et BacA-like sont essentiels pour le processus de différenciation, mais leur implication dans la symbiose avec les clades de légumineuses sous-étudiées, leur origine et leur histoire évolutive restent inconnues. Il a été montré précédemment qu'il existe cinq groupes homologues de BacA (BacA, BclA, ExsE, Mycobacterium BacA and Bradyrhizobium BacA), mais seulement deux d'entre eux sont impliqués dans la symbiose (BacA et BclA) (Guefrachi et al., 2015). Cependant, BacA et BclA sont des transporteurs de peptides localisés au niveau de la membrane interne, différant

principalement par la présence d'un domaine ATPase dans la protéine BclA qui est absent dans BacA. L'hydrolyse de l'ATP par le domaine ATPase est essentielle pour l'activité de transport de BclA, tandis que le transport par médiation de BacA est entraîné par la force proton motrice. Malgré cette différence, BacA et BclA peuvent importer des peptides NCR, et leur mutant de délétion rend les rhizobia hypersensibles à l'exposition de peptide NCR *in vitro.* Ainsi, la question motivant ce travail était de savoir si les familles de protéines BacA et BclA partagent une ascendance commune (par exemple, que BacA a évolué à partir de BclA, ou vice versa) ou si elles ont évolué indépendamment et ont convergé vers une fonction similaire. Ainsi, afin de résoudre l'histoire évolutive des transporteurs BacA, nous avons aussi combiné des analyses bioinformatiques (par ex. analyse phylogénétique des protéines BacA, analyse des séquences multi-locus pour reconstruire l'évolution du domaine des bactéries) avec des analyses fonctionnelles.

Tout d'abord, pour les analyses phylogénétiques, les souches de références des protéomes bactériens RefSeq avec un niveau d'assemblage complet ont été téléchargées à partir de NCBI. Nous avons conservé une espèce par genre. Ensuite, pour la phylogénie de BacA, les protéines BacA ont été recherchées dans les protéomes en utilisant hmmsearch (S. Eddy, 2009) avec l'alignement SbmA_BacA de Pfam (Finn et al., 2016). Les protéines trouvées ont été comparées à la base de données HMM des cinq homologues BacA (Guefrachi et al., 2015) en utilisant hmmscan (S. Eddy, 2009) et le HMM avec le meilleur score a été utilisé pour annoter chaque protéine. Les protéines annotées ont été alignées en utilisant le meilleur alignement évalué avec Tcoffee (Notredame et al., 2000) et cet alignement a été utilisé pour inférer la phylogénie des homologues de BacA en utilisant la méthode de maximum de vraisemblance sur IQtree2 (Minh et al., 2020). En revanche, pour la phylogénie des bactéries, les protéines conservées dans 95% des protéomes ont été identifiées en utilisant AMPHORA (Wu & Scott, 2012). Ces 32 protéines ont été concaténées et alignées pour ensuite être utilisé pour construire une phylogénie de maximum de vraisemblance du domaine des bactéries en utilisant aussi IQtree2 (Minh et al., 2020). En outre, les séquences homologues de BacA ont été utilisées pour construire un réseau de similarité des séquences (Oberg et al., 2023).

Cette analyse nous a permis d'identifier 366 homologues de BacA, dont 71 orthologues de BacA et 177 orthologues de BclA, à partir d'une recherche au sein des protéomes de 1255 espèces bactériennes. La première observation de nos phylogénies révèle que l'ordre des Hyphomicrobiales peut être subdivisé en deux groupes frères (*sister clades*) dans la phylogénie bactérienne ; BacA est largement distribué dans l'un de ces clades, tandis que BclA est

largement distribué dans l'autre. Cela pourrait suggérer que les protéines BacA et BclA de l'ordre des Hyphomicrobiales ont évolué à partir d'une protéine ancestrale commune présente dans l'ancêtre commun de ces clades. Cependant, les protéines Hyphomicrobiales BacA et BclA sont polyphylétiques dans la phylogénie des protéines BacA/BclA, ce qui suggère que les protéines BacA et BclA de l'ordre des Hyphomicrobiales ont été acquises indépendamment. Le regroupement distinct des protéines BclA et BacA dans le réseau de séquences soutient en outre des origines évolutives indépendantes pour ces protéines, tout comme la branche particulièrement longue reliant le clade BacA au reste de la phylogénie. De plus, nous considérons que les différences dans les mécanismes de transport de BacA (énergisé par gradient de protons) et BclA (énergisé par hydrolyse d'ATP) sont plus faciles à expliquer si ces familles de protéines ont des antécédents évolutifs distincts. En général, nous considérons que ces résultats suggèrent que les familles de protéines BacA et BclA ont évolué de manière indépendante, et que leur similarité fonctionnelle est le résultat d'une évolution moléculaire convergente.

Contrairement à nos attentes initiales, nous avons constaté que les deux familles de protéines présentent une distribution taxonomique limitée. Les orthologues BacA ont été identifiés uniquement dans le phylum Pseudomonadales, avec 89% des protéines BacA identifiées codées par des espèces des classes Alphaproteobacteria et Gammaproteobacteria. Une majorité des protéines BclA identifiées ont également été trouvées dans les espèces du phylum Pseudomonadales avec un biais vers les Betaproteobacteria. Cependant, les protéines BclA étaient également communes dans le phylum Cyanobacteriota, la classe Negativicutes (phylum Bacillota), et l'ordre Mycobacteriales (phylum Actinomycetota). L'observation que la plupart des clades taxonomiques enrichis pour des espèces codant pour BacA ou BclA contiennent également de nombreux organismes mutualistes et/ou pathogènes peut suggérer que l'interaction de l'hôte-eucaryote est un moteur du maintien de BacA et BclA dans ces lignées. Cependant, en supposant que BacA a été acquis par l'ancêtre commun de la sous-clade contenant BacA de l'ordre des Hyphomicrobiales, la famille de protéines BacA a potentiellement évolué dans cette lignée il y a plus de 500 millions d'années, événement qui précède l'évolution des légumineuses qui seraient apparues il y a environ 60 millions d'années. Donc, BacA n'aurait pas pu évoluer dans cette lignée en réponse à la symbiose des légumineuses. Nous émettons plutôt l'hypothèse que BacA a initialement évolué pour remplir un autre rôle et a ensuite été recruté pour la symbiose des légumineuses dans les rhizobia. De même, nous émettons l'hypothèse que BclA existait déjà dans la lignée de Bradyrhizobium

avant l'évolution de la symbiose avec les légumineuses, et que cette protéine a été recrutée indépendamment pour la symbiose des légumineuses dans ces organismes, imitant l'évolution convergente des peptides NCR.

Par la suite, des tests fonctionnels ont été effectués afin de tester la capacité de plusieurs protéines BclA récemment identifiées à complémenter le phénotype d'un mutant *S. meliloti* ΔbacA. Toutes les protéines étaient capables de complémenter au moins partiellement le phénotype de résistance à la gentamicine du mutant Δ*bacA*, suggérant que ces protéines étaient exprimées et au moins partiellement fonctionnelles chez *S. meliloti*. Cependant, comme prévu, seules les protéines annotées comme BclA étaient capables de complémenter efficacement la sensibilité au peptide NCR247 des mutants de *S. meliloti* Δ*bacA* et de *S. meliloti* Δ*bacA* Ω*yejA*.

3. Conclusion

En résumé, la combinaison des analyses d'évolution moléculaire basée sur la séquence et la structure avec des expériences fonctionnelles nous a permis d'acquérir de nouvelles connaissances sur l'évolution et la fonction des peptides NCR et des transports BacA.

Notre étude basée sur les séquences a révélé des peptides NCR clade-spécifique et espèce-spécifique, mettant en évidence la divergence des séquences d'acides aminés des peptides NCR connus, même à l'intérieur du clade et chez la même espèce. Cependant, l'analyse des peptides NCR dans un spectre plus large, en élargissant notre ensemble de données à quatre clades de légumineuses qui induisent la TBD et 3710 NCRs, nous a permis d'identifier des clusters de peptides NCR inter-clades. L'analyse structurale des peptides NCR a permis d'élucider d'importantes caractéristiques évolutives de la diversification des peptides NCR que nous n'avons pas pu résoudre avec des approches basées sur des séquences. Malgré la variation génomique des peptides NCR et des défensines, leurs structures 3D sont relativement conservées. En plus des changements du cycle cellulaire induits par les peptides NCR prédits, ces résultats suggèrent que les peptides NCR qui ont des séquences non apparentées mais des structures similaires et provenant d'espèces de légumineuses relativement éloignées (Dalbergioïdes-Indigoféroïdes) ont évolué à plusieurs reprises à partir du pool de défensines. D'autre part, les peptides NCR avec des séquences et des structures similaires des espèces IRLC et Génistoïdes relativement éloignées ont évolué à plusieurs reprises par évolution convergente, suivie d'une expansion probablement par duplication locale et d'une diversification rapide au cours de laquelle leur séquences divergent et leurs similitudes s'atténuent, menant à des

répertoires de peptides NCR spécifiques à l'espèce.

Concernant la distribution des transporteurs BacA dans le domaine des bactéries, leur évolution et leur fonction, nous avons identifié 208 espèces bactériennes portant au moins un gène codant pour BacA ou BclA. Ces espèces n'étaient pas également réparties dans le domaine des bactéries ; au lieu de cela, les protéines BacA ont été trouvées uniquement dans le phylum Pseudomonadota, tandis que les protéines BclA ont été principalement trouvées dans un sous-ensemble de familles à travers quatre phyla. Nos analyses suggèrent que les familles de protéines SbmA/BacA et BclA sont apparues indépendamment et que leur similarité fonctionnelle est le résultat d'une évolution convergente plutôt que d'une ascendance partagée. Nos données soutiennent également l'hypothèse que les protéines BacA et BclA ont été recrutées à plusieurs reprises pour faciliter les associations mutualistes et pathogènes avec les hôtes eucaryotes en permettant aux bactéries de faire face aux peptides antimicrobiens codés par l'hôte.

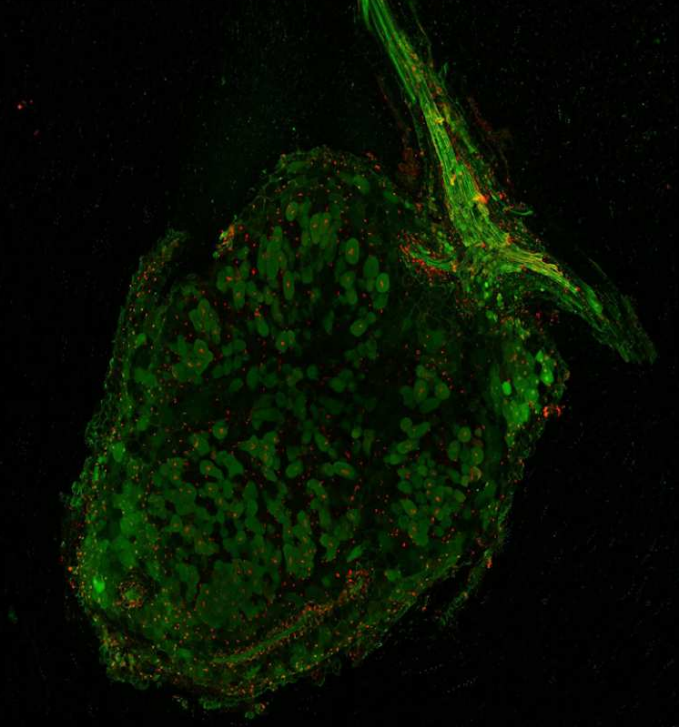
Pour conclure, cette étude nous a permis de déchiffrer l'histoire évolutive des peptides NCR d'un côté des transporteurs BacA de l'autre côté ou les deux types de protéines semblent être principalement recrutés par évolution convergente avec quelques cas particuliers et de formuler des hypothèses quant à la coévolution de ces deux protéines. Cependant, une étude fonctionnelle plus approfondie est nécessaire pour savoir si une coévolution convergente conduit l'évolution de ces deux protéines.

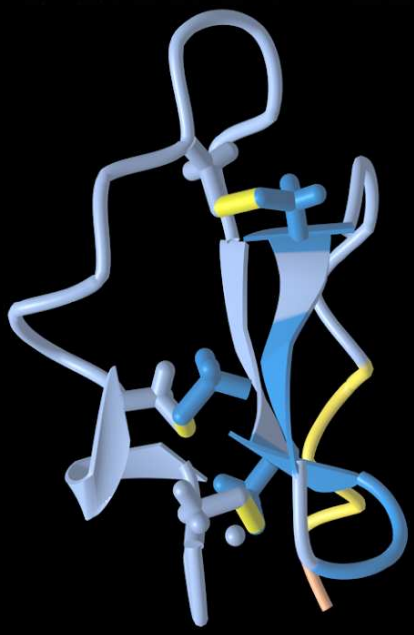
OUTLINE

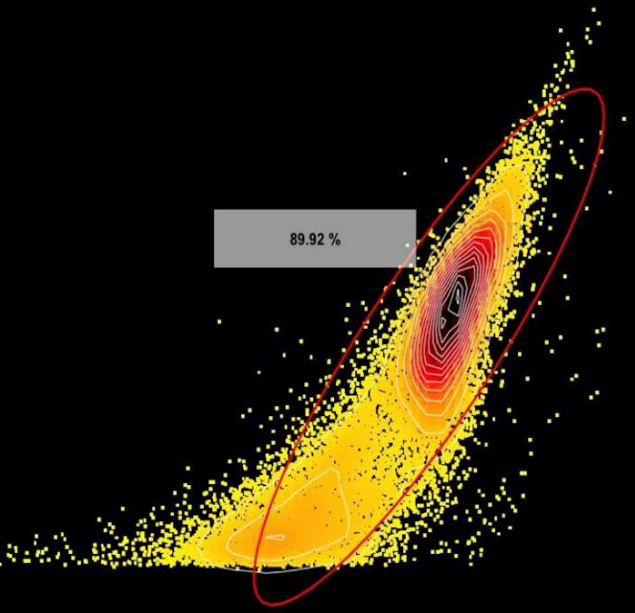
FIGURES

SUPPLEMENTARY FIGURES

SUPPLEMENTARY TABLES







I. Introduction

1. Concepts of evolution and coevolution

A. The concept of evolution

The concept of evolution belongs to the most important scientific theories. Charles Darwin postulated the theory of evolution by means of natural selection in the 19th century (Darwin, 1859). It provides a mechanism to explain how the remarkable diversity of living things has arisen from a common ancestor by natural processes and how they adapt to their environment.

Evolution is driven by many interconnected mechanisms, each of which influences and modifies life forms over time. During the 1800s, Charles Darwin and Alfred Russel Wallace proposed a concept of evolution by natural selection (Darwin, 1859)—a well-known theory elucidating how lifeforms adapt to their environmental and living surroundings. "On the Origin of Species" (Darwin, 1859) a book by Darwin, initially detailed this theory elaborately. Natural selection drives adaptation to particular environments because individuals who possess advantageous traits for their particular situation are more likely to establish, survive, and reproduce than those who do not have such traits (Brandon, 1978). This leads to the accumulation of beneficial traits within populations. The modern synthetic theory of evolution includes Darwin and Wallace's original mechanism of evolution by natural selection reconciled with theories of inheritance based on Mendel's findings (WELDON, 1902). The modern synthesis gave rise to the fields of population genetics and, more recently, population genomics. As we now accept, it is genetic information, not traits themselves, that is transmitted across generations (Mishra & Tatum, 1973). Different processes, including natural selection, random gene frequency fluctuations, known as genetic drift, migration leading to gene exchanges between populations, called gene flow, and unexpected changes impacting DNA sequence and arrangement, termed mutations, modify the frequencies of gene variants within populations and drive evolution, some of which is adaptive (Sniegowski & Lenski, 1995) **(Figure 1)**.

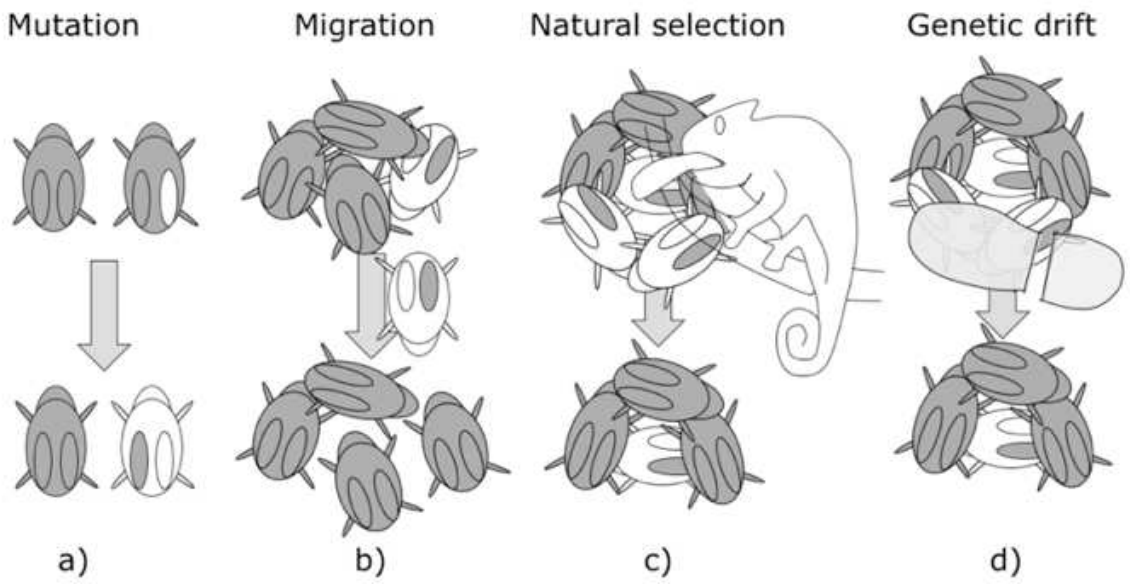


Figure 1 The different mechanisms of evolution.

Natural selection leads to adaptation to specific environments through differential survival and reproduction of those individuals that fit it best. Migration and mutation are random events that introduce new variations, and genetic drift generates random fluctuations in trait frequency between generations (Schlieter et al., 2019).

Convergent evolution can be observed when species that do not share a direct ancestor independently evolve comparable characteristics that deal with similar environmental challenges (Doolittle, 1994). For example, wings in insects and bats enable flight but have different anatomical origins. On the other hand, when related species with a recent common ancestor share similar derived phenotypes **(Figure 2bc)**, we talk about parallel evolution (Zakon, 2002). Divergent evolution arises when a single species or population branches and diverges into two or more distinct, differentiated populations or species over time. Another important evolutionary pattern is coevolution, which occurs when distinct organisms mutually influence each other's evolutionary history (Ehrlich & Raven, 1964). This kind of evolutionary pattern is the basis of a lot of parasitic and mutualistic symbioses.

Furthermore, there are many areas where the theory of evolution has significant implications, including medicine and biodiversity conservation (Nesse et al., 2006). Evolutionary principles guide conservation biology strategies aimed at conserving genetic diversity. Studying the genetic basis of diseases and the development of antibiotic resistance is needed to develop beneficial treatments and address issues related to public health. Enhancing nitrogen fixation symbiosis to lower fertilizer use and breeding crops resistant to pests are guided by evolutionary insights in agriculture.

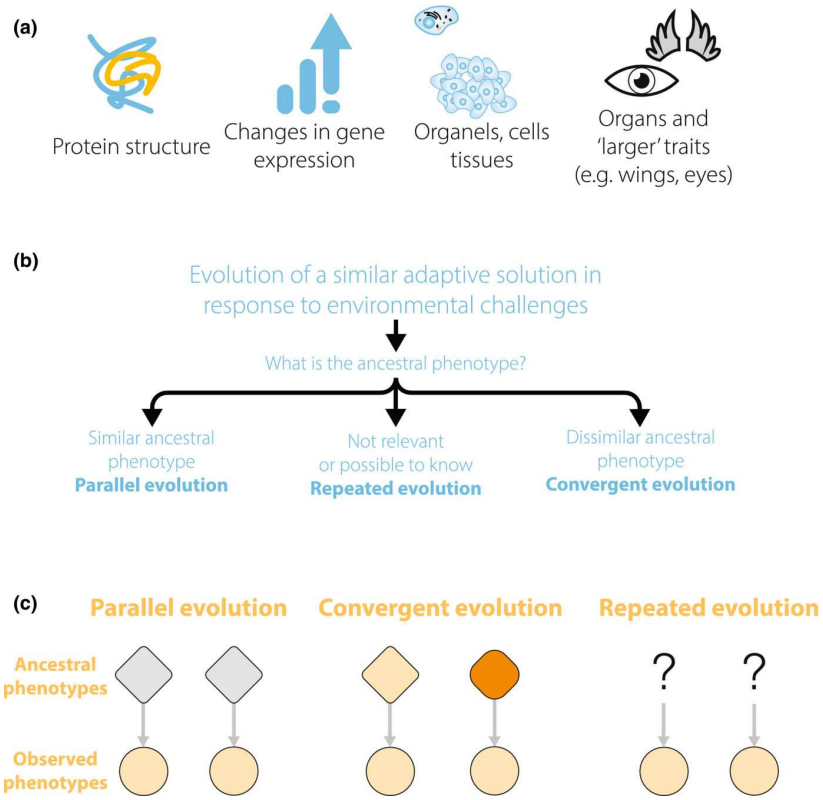


Figure 2 The different evolutionary patterns in response to the environmental changes.

(a) There are different levels where natural selection can operate at the phenotypic level. (b) The various possibilities of evolution in response to similar environmental pressures. (c) Representation of parallel, convergent, and repeated evolution and their differences (Cerca, 2023).

B. Is evolution repeatable?

"Related species will vary in similar directions and be subject to similar selective influences. They may, therefore, be expected to evolve in parallel." Haldane, *The causes of evolution,* 1932, p. 76-77 (Gompel & Prud'homme, 2009).

As described above, evolution is how living beings transform, evolve over time, and are selected by their environments based on the principles of natural selection. Evolution's repeatability remains a challenging question in evolutionary biology, an evolutionary topic that has engaged biologists in lengthy debates. In other words, would we see similar evolutionary outcomes if we could replay life's tape several times? This idea was famously encapsulated in Stephen Jay Gould's thought experiment of « replaying the tape of life » to see if the same events would unfold.

The debate was about whether evolution was more deterministic versus contingent (non-deterministic). Stephen J. Gould, an anti-adaptationist, proposed that if the history of life on Earth could be rewound and played again from the beginning, the outcomes would be vastly different due to the contingency in evolution. This notion highlights the significance of random events, historical contingencies, and the complex interactions between organisms and their environments in shaping the course of evolution while underplaying the role of adaptation (Blount et al., 2018). Gould's argument emphasizes that small changes at critical decision points in evolutionary history could lead to different outcomes, illustrating the non-repeatable nature of evolution. By introducing the concept of contingency, Gould questioned the predictability and repeatability of the evolutionary trajectories, suggesting that non-deterministic (unpredictable) factors largely influence evolution. Nonetheless, most evolutionary biologists agree that adaptation is a fundamental driver of both contingent and repeatable evolution. While chance and historical contingencies certainly also influence evolutionary trajectories, the repeated evolution of traits across different lineages underscores the importance of adaptation in shaping organisms in response to environmental pressures.

Some studies (Lind, 2019) show that evolution is repeatable at some scales, i.e., different populations or different related species frequently acquire similar phenotypes independently. For example, bacteria become resistant to antibiotic drugs, multiple human populations independently evolved the retention of the ability to digest lactose at an advanced age, birds living on islands have independently lost flying abilities, and different insects have independently evolved dark-colored marks on their wings: all cases of parallel evolution with similar traits evolving independently across distinct lineages. **(Figure 3).**

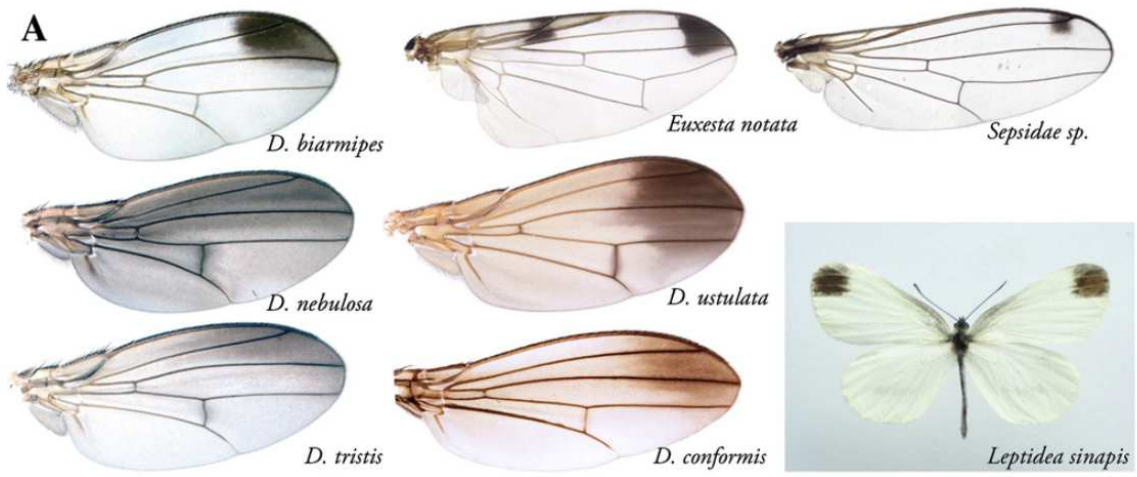


Figure 3 The repeated evolution of similar dark-wing pigmentation independently in different fly species. (Gompel & Prud'homme, 2009).

Conversely, numerous studies have illustrated that evolution might greatly depend on contingency. This signifies that the results of evolutionary processes hinge on particular historical and environmental conditions. In laboratory settings, "parallel replay" experiments, like the Long-Term Evolution Experiment with *E. coli* (LTEE), reveal that identical populations under the same conditions can evolve differently, illustrating contingency in evolution (Blount et al., 2018). Moreover, studies like the comparison of woodpeckers and aye-ayes adapted to the same feeding niche but evolving different traits to access grubs in the trunks of trees underscore how evolutionary legacies affect current adaptations, showcasing the contingent nature of evolution (Blount et al., 2018). Clearly, adaptation and contingency play a role in character evolution. For example, it has been recently suggested that genome evolution in endosymbionts bacteria is both deterministic, favoring B-vitamin genes, and stochastic, leading to diverse gene inventories with limited redundancy (Boyd et al., 2024).

Various characteristics impact the repeatability of evolution, including genetic relatedness. Replicate populations that are closely related genetically tend to show more repeatability in evolutionary outcomes, possibly due to shared genetics and developmental pathways (Blount et al., 2018). Additionally, evolutionary convergence is more likely among lineages sharing similar natural environments, highlighting the impact of environmental factors on repeatability. This demonstrates the power of natural selection in sculpting similar adaptive solutions repeatedly. The presence of historical differences among populations can reduce the likelihood of repeatable outcomes in evolution. Divergence in evolutionary histories can lead to variation in adaptive responses to similar conditions, emphasizing the role of contingency and historical

influences (Blount et al., 2018). Furthermore, chance events, such as specific mutations, can introduce variability in evolutionary trajectories, affecting the repeatability of outcomes in evolution. Even starting from identical conditions, different mutational occurrences can result in divergent evolutionary paths (Blount et al., 2018). Therefore, evolution's contingency and repeatability are not universal but dependent on the context and the factors used to define it and can occur simultaneously.

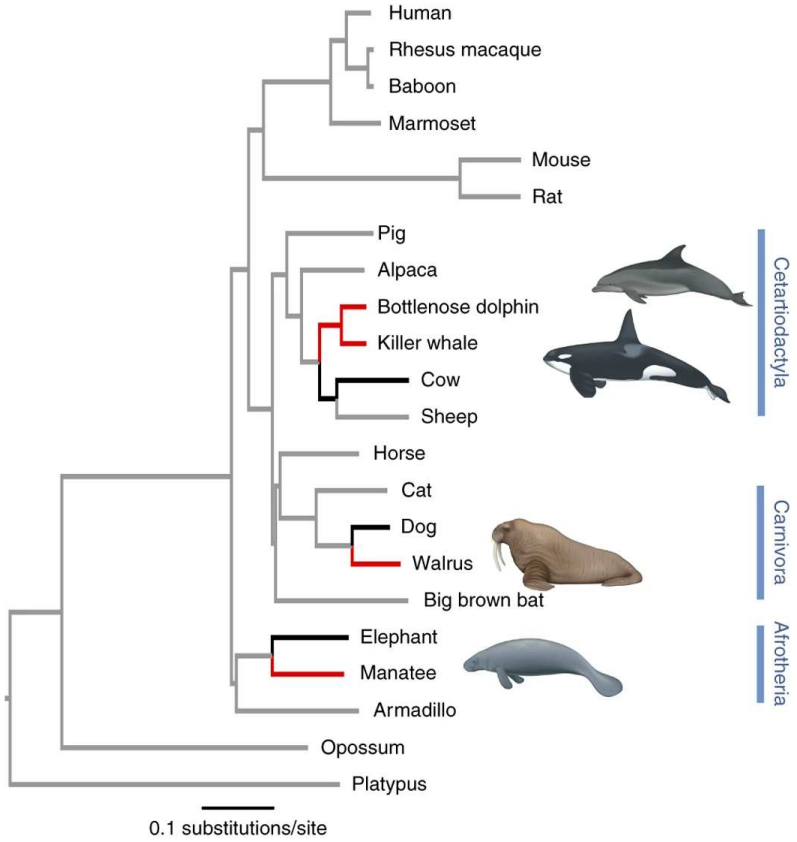


Figure 4 Phylogeny of 20 eutherian mammalian genome sequences, rooted with a marsupial outgroup. The branches colored in red represent the independent evolution of marine mammal lineages, for which tests for positive selection and parallel nonsynonymous amino acid substitutions were performed. Branches of the control set of terrestrial taxa, for which tests for positive selection and parallel nonsynonymous amino acid substitutions were also performed, are colored black (Foote et al., 2015).

When similar traits evolve repeatedly in different organisms in independent lineages facing the same environmental conditions, we consider this evidence that the traits are adaptations, evolving via natural selection. For example, marine mammals that evolved from three distinct mammalian lineages (**Figure 4**) share a number of very similar derived traits despite their independent ancestry, strongly suggesting that these traits are adaptations to the marine lifestyle (Foote et al., 2015; Reidenberg, 2007). These similar phenotypes, in response to similar environmental challenges, may or may not share similar molecular genetic mechanisms. While

various case studies (Gompel & Prud'homme, 2009) indicate that repeated phenotypic evolution can result from similar genetic changes, the study about marine mammals (Foote et al., 2015) suggested that whereas molecular convergent evolution is common, molecular convergence exclusively linked to phenotypic convergence is relatively rare. However, only a few studies have addressed this question, and we still wonder if phenotypic similarities can arise from different mechanisms or have evolved independently using similar genetic mechanisms.

C. Convergent evolution and coevolution

Convergent evolution is the repeated appearance of the same or similar character states in independent lineages. Distant species may display similar features separately, usually in response to comparable situations or habits., as shown in the marine mammal example above. These traits illustrate parallel responses to environmental challenges. Examples of convergent evolution are observed across various biological systems, from plants to mammals (Foote et al., 2015). Studies have shown that convergent evolution can occur at different levels, including whole organisms, organ systems, gene networks, and specific proteins (Brazhnik & Tyson, 2006). It is not limited to specific taxa but is widespread in nature (Arbuckle et al., 2014). The convergence of phenotypic traits can occur through both parallel and nonparallel mechanisms, challenging the assumption that similar phenotypes among closely related species evolve through the same mechanisms (Twomey et al., 2023). Furthermore, the study of molecular convergence using genomic data not only reveals the molecular basis of phenotypic convergence but also provides insights into the principles of convergent evolution (S. Xu et al., 2020).

One classic example of convergent evolution is the thylacine and canids, which exhibit striking similarities despite their ancient divergence time, making them a widely recognized case of convergent evolution (Feigin et al., 2018). Another example is the convergence in toxin resistance across different animal species due to similar selection pressures, leading to predictable evolutionary responses (Arbuckle et al., 2017; Ujvari et al., 2015).

Convergent evolution provides biologists with an invaluable chance to investigate the processes of adaptation, selection, and the intricate relationships between species and their habitats. By utilizing interdisciplinary approaches and examining a wide range of species, scientists may enhance their comprehension of evolutionary convergence and its significant influence on the

variety of living forms' evolution. In summary, convergent evolution serves as a cornerstone in understanding the repeatability of evolution by demonstrating how similar traits can independently arise in different lineages, shedding light on the predictability and underlying genetic mechanisms driving evolutionary processes.

As observed in different species interactions, coevolution is an essential example of evolutionary repeatability across ecosystems (Agrawal & Zhang, 2021). We talk about coevolution, when two or more interacting species reciprocally influence each other's evolutionary histories through natural selection. For example, the arms race between predators and prey often leads to adaptations in both, such as enhanced velocity (Nair et al., 2019).

Coevolution, as exemplified by Darwin's famous hawk moth and orchid interaction, showcases the reciprocal evolution of interacting species (**Figure 5**). This mutualistic relationship between the hawk moth and the orchid highlights the specialized adaptations that have evolved over generations. Darwin's observations emphasized the intricate fit between the nectar spur of the orchid and the proboscis's length of the pollinator, illustrating the concept of coevolution driving trait evolution in both partners (Harder & Johnson, 2009).

Figure 5 The Darwin's hawk moth coevolution.

The hawk moth *Xanthopan morganii praedicta* pollinates the Madagascar star orchid (*Angraecum sesquipedale*) with its long proboscis (source: **Encyclopedia Britannica**).

According to Ehrlich and Raven (Ehrlich & Raven, 1964), coevolution includes any ecological interaction that results in coadaptation among interacting species (competition, mutualisms, or antagonisms) **(Figure 6)** (Carmona et al., 2015). These dynamics result from selective pressures imposed by ecological interactions, driving species to adapt and counter-adapt over time. Ehrlich and Raven's work on coevolution provides the foundation for understanding the intricate relationships between species, particularly in the context of plant-insect interactions. They introduced the concept of stepwise coevolution, highlighting the reciprocal adaptations that occur between butterflies and angiosperms, leading to increased biological diversity within these groups (Wheat et al., 2007). The impact of plant-insect interactions on speciation has been a central theme in Ehrlich and Raven's work. They proposed that herbivorous insects have played a significant role in driving plant speciation, emphasizing the role of adaptive responses and speciation patterns influenced by interacting species (Ehrlich & Raven, 1964). Additionally, their coevolutionary theory predicted that the evolutionary success of entire insect and plant clades is governed by their reciprocal adaptations, supporting the idea of enemy-driven adaptive radiation through the evolution of plant defenses (Ehrlich & Raven, 1964).

Coevolution is a phenomenon that can be observed at both the phenotypic and molecular levels (**Figure 6**). While species-level coevolution occurs via reciprocal adaptations between interacting organisms, molecular coevolution is driven by changes in genomic sequences (**Figure 6**) (Carmona et al., 2015). At the phenotypic level, coevolution manifests in various ways, such as resistance, virulence, life-history trade-offs, and biodiversity (Brockhurst et al., 2004; Buckling & Rainey, 2002a, 2002b; Forde et al., 2008; Lohse et al., 2006). Detecting coevolution at the molecular level can be challenging due to the complexity of identifying the genes involved and the statistical complexities in determining coevolution (Codoñer & Fares, 2008; de Juan et al., 2013).

In summary, coevolution, as elucidated by Darwin, Ehrlich, and Raven and further investigated, highlights the interconnected nature of species interactions and evolutionary processes. From specialized plant-pollinator relationships to the emergence of diverse ecological networks, coevolutionary mechanisms play a pivotal role in shaping biodiversity and steering evolutionary diversification.

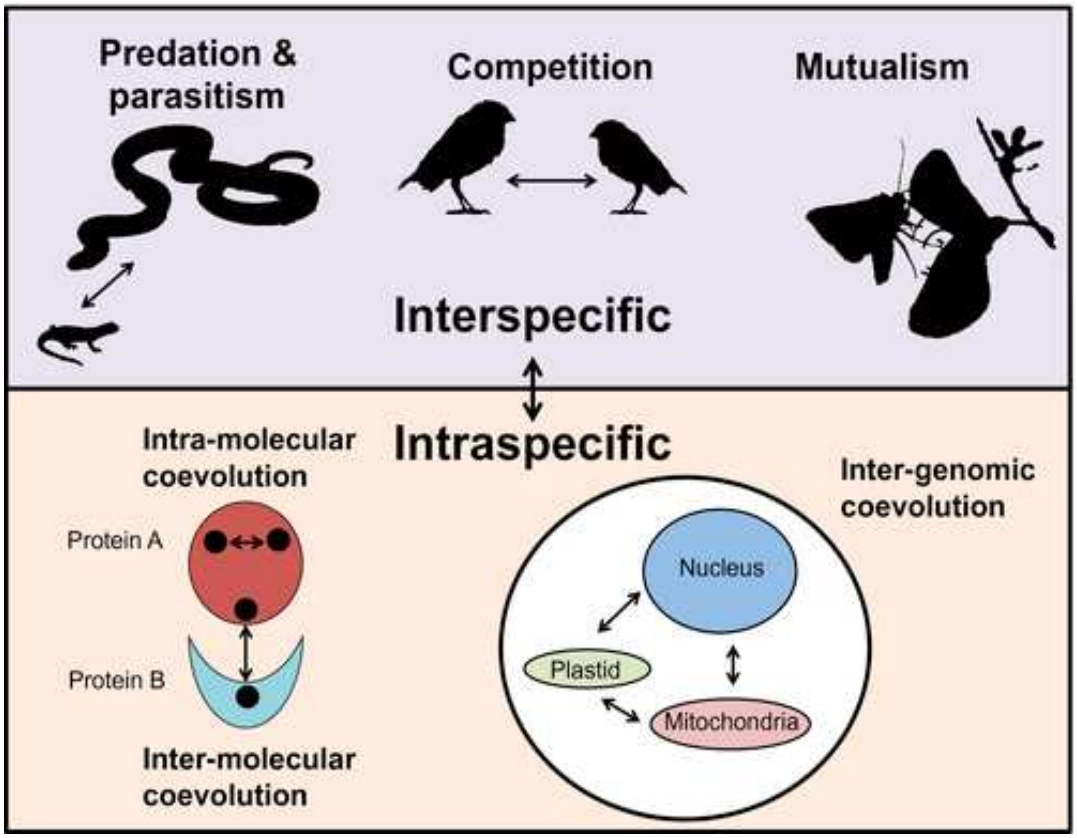


Figure 6 Co-evolution occurs at multiple levels of biological organization.

Co-evolution could be interspecific between different species and intraspecific between different proteins or organelles (Carmona et al., 2015).

Convergent coevolution occurs when unrelated pairs of interacting species confronted with similar ecological limitations independently evolve similar properties. For instance, similar floral traits, like nectar spurs, that guide and manipulate pollinating insects may evolve in independent lineages of flowering plants, driving and being driven by similar modifications of insect mouthparts (Hodges, 1997).

To comprehensively study convergent coevolution, a combination of molecular evolution analysis and functional experiments is essential. Molecular evolution analysis, often conducted through bioinformatics approaches, can reveal the genetic basis of convergent evolution (Sackton et al., 2019; Thomas & Hahn, 2015). By examining genomic data, we can identify convergent molecular changes in proteins or regulatory regions that drive phenotypic convergence (Sackton et al., 2019). Functional experiments, on the other hand, allow us to directly test the functional significance of these molecular changes and how they contribute to convergent phenotypic outcomes (Chirat et al., 2013; Nagy et al., 2014).

Furthermore, studying molecular coevolution involves analyzing the interactions and evolutionary dynamics between molecules, such as proteins or RNA, to understand how changes in one molecule influence changes in another over time. Various approaches and tools have been developed to investigate molecular coevolution, ranging from experimental evolution studies to computational methods. Bioinformatics provides a powerful tool for analyzing molecular sequences and structures to understand evolutionary relationships and predict functional implications (Ndagi et al., 2020). Functional assays play a vital role in complementing molecular evolution analysis by providing a more biologically relevant perspective (Han et al., 2019). For instance, in the study of Hepatitis C Virus fusion, combining coevolution analysis with in vitro assays uncovered functionally significant coevolving signals between specific regions governing the fusion process (Douam et al., 2018). By integrating bioinformatics with experimental techniques, we can detect and characterize coevolution within protein complexes without the need to identify specific correlated mutations in complex subunits (Sandler et al., 2013). In summary, by combining experimental evolution, convergence analysis, predictive structural bioinformatics, and omics approaches, we can gain insights into molecular convergence and coevolution, shedding light on the shared genetic changes that lead to similar phenotypic outcomes of unrelated pairs of interacting species.

Finally, convergent coevolution highlights the remarkable repeatability of evolution, showcasing how distinct lineages can undergo analogous evolutionary changes in response to similar selection pressures. These examples highlight the significance of ecological interactions in propelling evolutionary processes and offer strong evidence for the function of natural selection in forming biological diversity. Furthermore, investigations into coevolution's molecular and genetic foundations shed additional light on the mechanisms behind these repeating evolutionary patterns. So, coevolution and convergent coevolution point out the predictability and recurrence integrated into the evolutionary process by providing fascinating instances of how evolution leads to similar results confronting comparable environmental obstacles. However, while various studies have delved into different aspects of molecular coevolution, the availability of suitable models remains limited. Here, we use the legume-rhizobia mutualistic symbiosis model, an ecologically important model to study convergent coevolution.

2. Legume-rhizobia symbiosis – an example to study convergent coevolution

The legume-rhizobia symbiosis stands as a notable example of studying convergent coevolution, showcasing how different species can independently arrive at similar solutions through convergent evolution.

A. The concept of symbiosis

Symbiosis is an intriguing and crucial area in biological science that describes the tight and generally long-term relations between two organisms. Such interactions emerged over millions of years and may be beneficial or harmful, each of them exhibiting particular features and impacts on the species involved.

The term symbiosis, derived from the Greek words 'syn' meaning 'together' and 'biosis' meaning 'living', was proposed in 1879 by the German mycologist Heinrich Anton de Bary as "the living together of unlike organisms" when he studied lichens, which consist of a fungus living in symbiosis with algae or cyanobacteria (Raina et al., 2018). This broad term has evolved over time to include durable and close relationships that differ in their dependency and benefit to the species involved. The research on the topic of symbiosis has demonstrated its critical significance in the evolution and operation of many ecosystems.

We can distinguish mutualism, commensalism, and parasitism among these symbiotic interactions, with beneficial, neutral, or harmful outcomes to one or both interacting species (**Figure 7**). Among the most renowned examples of symbiotic mutualism, a relationship where both entities coexist for shared advantages, is the interaction between rhizobia bacteria and leguminous plants (**Figure 7A**) (Hirsch et al., 2001; E. T. Wang, 2019). Commensalism involves one organism benefiting while the other is neither helped nor harmed. An example would be the *Riptortus pedestris* gut symbiosis, where the bean bug *Riptortus pedestris* harbors *Burkholderia* spp. symbionts in its midgut (**Figure 7B**). The *Riptortus pedestris* gut symbiosis represents an adaptation to a diet that lacks certain nutrients. The bean bug has evolved a specialized gut region to house the *Burkholderia* symbionts, which help in nutrient supplementation. This relationship has likely evolved through a series of adaptations, allowing the bug to utilize the bacteria without harm (Lee et al., 2024). In a parasitic relationship, one entity gains advantages to the detriment of another. Pathogens, including *Mycobacterium tuberculosis* among mycobacteria, inhabit the host's cells and lead to diseases such as tuberculosis. They have evolved sophisticated mechanisms to evade the host’s immune system

and sustain themselves within the host (R. M. Jones & Neish, 2011).

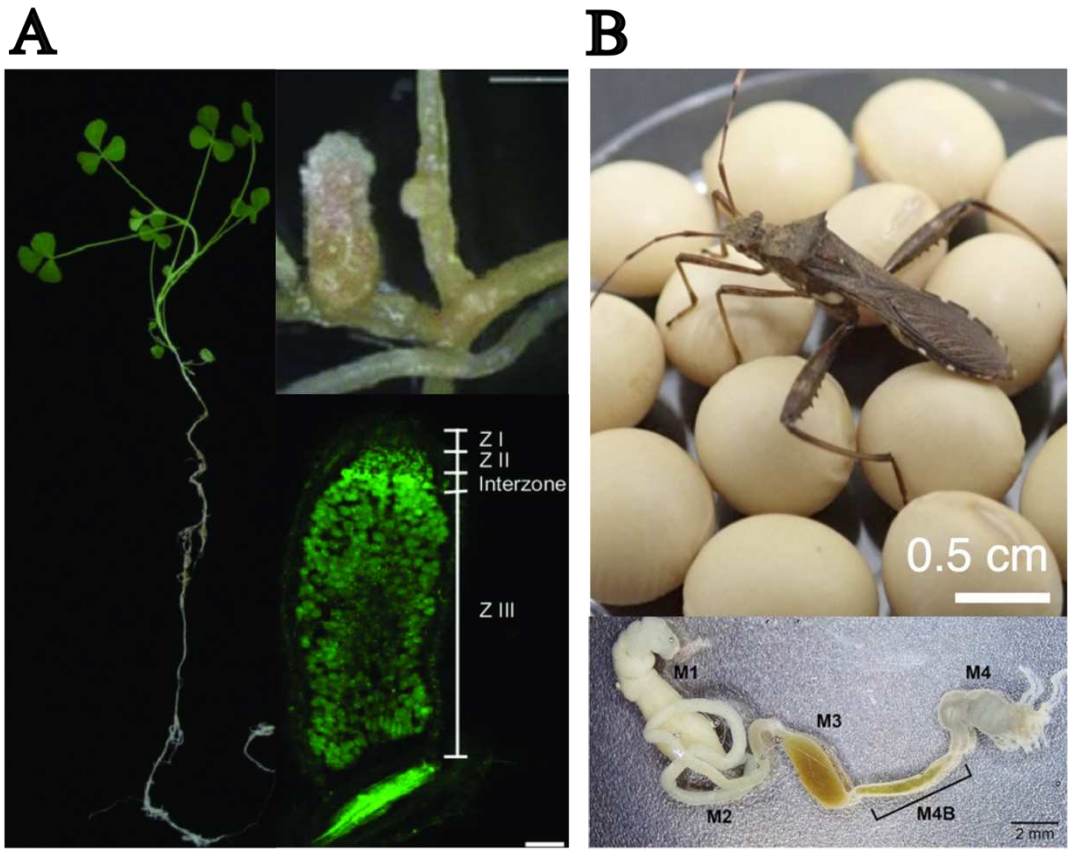


Figure 7 Examples of mutualism and commensalism symbiotic programs.

(A) The mutualistic nitrogen-fixing symbiosis between *Medicago truncatula* and *Sinorhizobium medicae* was adapted from (Walton et al., 2020). On the left, the *Medicago truncatula* 31 days post-inoculation. On the top right are the root nodules from the plants. On the bottom right is the confocal microscopy image of a longitudinal section of a root nodule stained with Syto9, highlighting intracellular symbiotic bacteria in green. (B) The bean bug *Riptortus pedestris* midgut symbiosis. On the top is an adult male of *R. pedestris* feeding on soybean seeds (Jang et al., 2021), and the dissected gut displays midgut regions (M4B) filled with the symbiotic bacteria Burkholderia insecticola (Lee et al., 2024).

Furthermore, endosymbiosis (Archibald, 2015) is involved in several major biological transitions, with the fusion of two prokaryotic cells giving rise to eukaryotes (W. F. Martin et al., 2015), and with the incorporation of cyanobacteria into a non-photosynthetic protist giving rise to photosynthetic eukaryotic organisms (Bhattacharya et al., 2004).

In summary, symbiosis is a crucial ecological concept that demonstrates the interdependence of species and shapes biodiversity. The legume-rhizobia mutualistic relationship is an excellent illustration of how symbiosis may be used for ecological and agricultural purposes.

B. The nitrogen-fixing symbiosis

Nitrogen is an essential element for life, a component of many molecules, such as nucleotides. Although nitrogen is the most abundant gas in the atmosphere, most of it is dinitrogen (N_2), a form not usable by most organisms due to the highly stable triple bound between the two nitrogen atoms. Only bacteria are capable of converting dinitrogen to a biologically available form via the process of Biological Nitrogen Fixation (BNF), which is essential for making nitrogen available for biological processes (Rosca et al., 2009). Biological Nitrogen Fixation consists of converting atmospheric nitrogen (N_2) into ammonia (NH_3), a form that organisms can use to produce their vital molecules. This process is essential for sustaining ecosystems by providing bioavailable nitrogen, which is necessary for plant growth and influences global elemental cycles (Yu & Zhuang, 2020). Through available nitrogen can be remobilized and recycled from dead organic matter, BNF is particularly important in nitrogen-limited ecosystems, such as dryland and Mediterranean regions, where it shapes the nitrogen-carbon cycle (Dovrat & Sheffer, 2019). It contributes to the nitrogen economy of aquatic ecosystems, although its importance varies depending on the ecosystem (Howarth et al., 1988). In addition, nitrogen fixation by cyanobacteria in constructed semi-aquatic ecosystems impacts nitrogen removal efficiency, demonstrating the significance of this process in engineered ecosystems as well (X. Zhang et al., 2017). Furthermore, in peatland ecosystems, nitrogen fixation by lichens has been identified as a contributing factor to the nitrogen balance in these environments (Waughman & Bellamy, 1980).

In addition to BNF, the artificial synthesis of fertilizers is also widely used in agriculture to provide plants with usable nitrogen. However, the synthesis of fertilizers has different disadvantages, such as water requirements and high greenhouse gas emissions, causing environmental pollution and health issues. For instance, the Haber-Bosch process (Haber & van Oordt, 1905), which involves the conversion of nitrogen and hydrogen into ammonia, has been pivotal in agricultural practices and is credited with significantly increasing food production globally (Smil, 1999). This process is responsible for producing over 90% of the world's ammonia. It has been instrumental in feeding around 40% of the world's population (Cherkasov et al., 2015). Despite its importance, the Haber-Bosch process is energy-intensive and poses environmental challenges due to its reliance on fossil fuels and high temperatures and pressures (L. Wang et al., 2018). The traditional method has been criticized for its significant energy consumption, with estimates suggesting that it consumes about 1% of the total energy production globally and contributes to approximately 1.4% of global CO_2 emissions (Zhu et al.,

2022). Therefore, using BNF instead of synthetic fertilizers for nitrogen supply is crucial.

Biological nitrogen fixation is achieved by some prokaryotes known as diazotrophs using nitrogenase enzymes that convert atmospheric nitrogen N_2 into ammonia NH_3, which can be used by plants and other organisms (Zehr et al., 2001). Symbioses between nitrogen-fixing prokaryotes and eukaryotes are crucial for nitrogen acquisition in nitrogen-poor environments (Thompson et al., 2012). Nitrogen-fixing symbioses play a crucial role in the plant kingdom, particularly in legumes, where root nodules are formed in association with rhizobial bacteria. These root nodules are a result of a complex signal exchange mechanism between the bacteria and the plant, leading to the formation of invasion structures for bacterial entry into the plant root (K. M. Jones et al., 2007). Additionally, actinorhizal plants (a polyphyletic group including some Rosales, Cucurbitales, and Fagales) engage in nitrogen-fixing symbiosis with several species of the bacterial genus *Frankia*, thus several unrelated groups of plants outside of legumes (Fabales) have nitrogen-fixing capacity (Ourèye Sy et al., 2007; Sen et al., 2013). While most nitrogen-fixing symbioses involve root nodules, there are exceptions, like the nitrogen-fixing gram-negative bacterium *Azospirillum*, which fixes nitrogen without forming root nodules (Mehnaz, 2011). In specific plant-microbe interactions, such as with the water mimosa plant *Neptunia natans*, a nitrogen-fixing root-nodule symbiosis that forms only one nodule, showcasing the diversity of symbiotic relationships in the plant kingdom (Rivas et al., 2002). Moreover, *N. natans* engages in a nitrogen-fixing root-nodule symbiosis with two distantly related bacteria, *Devosia neptuniae* and *Allorhizobium undicola* (De Lajudie et al., 1998; Rivas et al., 2003).

The most important nitrogen-fixing symbiosis that can replace use of artificial fertilizers is the classic mutualism between legume plants and rhizobia bacteria. The umbrella name “rhizobia” includes alpha and beta proteobacteria (Willems, 2006). Many legumes can form a nitrogen-fixing symbiosis with rhizobia. During this interaction, the legume plant forms root nodules where the bacteria are housed intracellularly inside structures called symbiosomes where the rhizobia fix atmospheric nitrogen and transfer ammonia to the legume plant using the nitrogenase enzymatic complex comprising NifDK and the dinitrogenase reductase NifH. In return, the plant provides carbon to the bacteria.

The evolutionary history of nitrogen-fixing rhizobia bacteria reveals that the common ancestor of rhizobia did not possess symbiotic genes and that the ability to fix atmospheric nitrogen in nodules was gained multiple times independently (Garrido-Oter et al., 2018) (**Figure 8)**.

Nitrogen-fixing ability was probably acquired initially from non-rhizobial species, possibly by horizontal gene transfer, which can also explain the polyphyletic origin of rhizobia. On the other hand, the nitrogen fixation ability in plants was gained several times in different clades (**Figure 9**) (Delaux et al., 2015). We also know that the co-evolution of legume plants with rhizobia bacteria has influenced the evolution of legumes, which have evolved different features to co-exist and promote the survival of their rhizobial bacteria (Martínez-Romero, 2009).

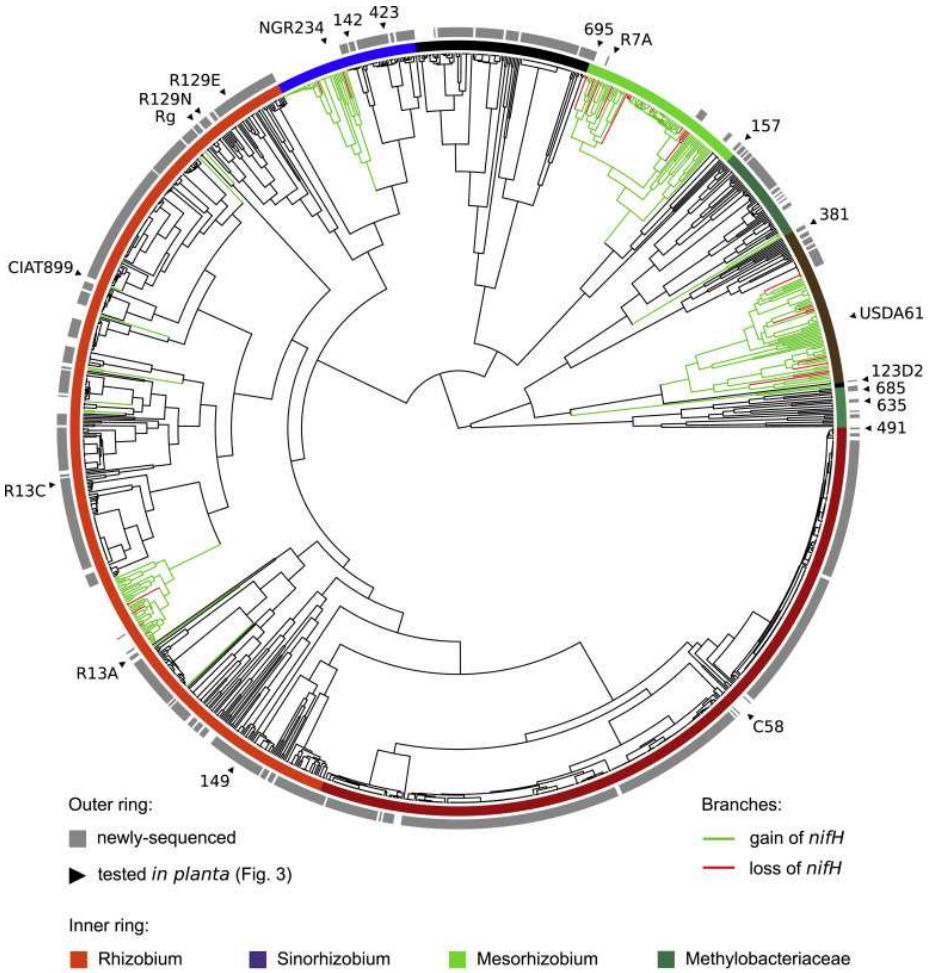


Figure 8 Maximum likelihood phylogenetic tree of whole genome single-copy symbiotic genes displaying an ancestral state reconstruction for *nifH* presence.

This tree reveals that these genes were gained independently multiple times from a non-rhizobial species by horizontal transfer (Garrido-Oter et al., 2018).

In summary, the nitrogen-fixing symbiosis between legume plants and rhizobia bacteria is important for promoting nitrogen cycling and plant growth. Therefore, understanding the mechanisms underlying this process and the evolutionary history of this symbiosis is important to enhance sustainable agriculture.

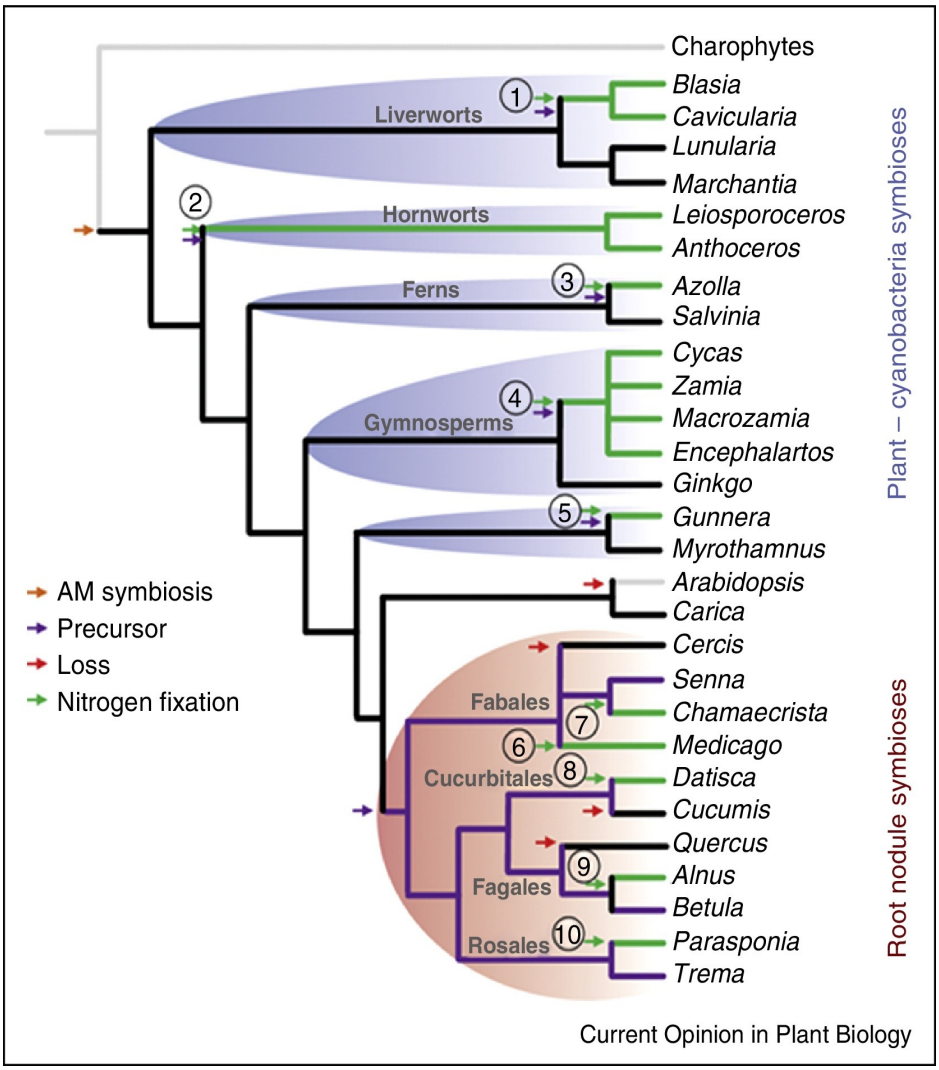


Figure 9 Evolution of nitrogen-fixing symbiosis in plants.

The gain of nitrogen-fixation ability is represented by the green arrow in the tree. The gray branch represents a clade that forms no symbiosis. The black branches represent the branches that form mycorrhizal symbiosis. The purple branches are for predisposed clades, and the green clades are for nitrogen-fixing species clades (Delaux et al., 2015).

C. Legume plants

Legumes are a diverse group of plants that contain economically important grain legumes, oilseed crops, forage crops, shrubs, and trees (Singh et al., 2007). They are an exclusive group of plants known for their ability to perform a symbiotic interaction with nitrogen-fixing bacteria (A. Liu et al., 2020) that fix atmospheric nitrogen. This feature enables them to grow in nitrogen-poor soils and to provide nitrogen for other plants, rendering them important for improving environmental quality and agriculture (Ritchie & Tilman, 1995). This symbiosis allows the legumes to fix atmospheric nitrogen, enriching the soil with nitrogen and reducing

the need for synthetic fertilizer.

Legumes, with their wealth of available nitrogen, even when growing on poor soils, are protein-rich, making them an essential source of food (Amarowicz, 2020; Voisin et al., 2014). They represent one of the most important food sources after cereals (Varshney & Kudapa, 2013).

The *Fabaceae* family, also known as *Leguminosae,* is the third-largest Angiosperm family of flowering plants, containing more than 19,500 species regrouped into 765 genera and six sub-families (Azani et al., 2017). The *Papilionoideae* is the largest sub-family of legumes (**Figure 10)**, with more than 14,000 species, 501 genera, and 32 tribes (Silva et al., 2022). In terms of agriculture and economy, this family is important because it includes the most important cultivated legumes, such as peanut (*Arachis hypogaea*) and bean (*Phaseolus vulgaris*), and forage crops like clover (*Trifolium pratense*) and alfalfa (*Medicago sativa*). The majority of the *Papilionoideae* sub-family species can enter a symbiotic interaction with bacteria to fix atmospheric nitrogen.

The *Papilionoideae* sub-family is split into six different clades, namely Genistoids, Dalbergioids, Robinioids, Millettioids, Indigoferoids, and the inverted repeat-lacking clade (IRLC) (**Figure 10**). The IRLC (Inverted Repeat Lacking Clade) is the largest monophyletic clade. This clade is characterized by the absence of a 25-kb inverted region in the chloroplast genome (**Figure 10**) (Choi et al., 2022). This clade includes the *Medicago* genus, the best-studied genus in legumes, and other important genera, such as *Pisum*, *Trifolium*, and *Astragalus*.

The Dalbergioid clade diverged from the other legume clades around 55 million years ago (**Figure 10**) (Lavin et al., 2005). The Millettioids clade includes economically important crops like soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*). The *Lupinus* genus, another study model in legume-rhizobia symbiosis, belongs to the Genistoids clade (**Figure 10**). The Robinoids clade consists of various genera, including the model legume *Lotus japonicus*. The Indigoferoids clade, which includes the *Indigofera* genus, is closely related to the Millettioids clade and split from the larger Indigoferoid/Millettioid clade. The genus *Indigofera*, known for its significant diversity with approximately 750 species, is the third-largest genus in the legume family (Schrire et al., 2009).

These six clades differ one from the other in their symbiotic characteristics and features, such as forming different types of nodules, managing bacteria differently, and varying in symbiotic efficiency.

Only a few species of the diverse *Papilionoideae* have been sequenced, genetically analyzed, and well-characterized at the molecular level. Among them are the model legumes *Medicago truncatula* and *Lotus japonicus* (Sato et al., 2008; Young et al., 2011). Other legume species have been sequenced but not studied enough at molecular levels, such as peanut (*Arachis hypogaea*). Furthermore, there are some *Papilionoideae* clades for which no species have been well-studied, such as the Genistoids and the Indigoferoids clades.

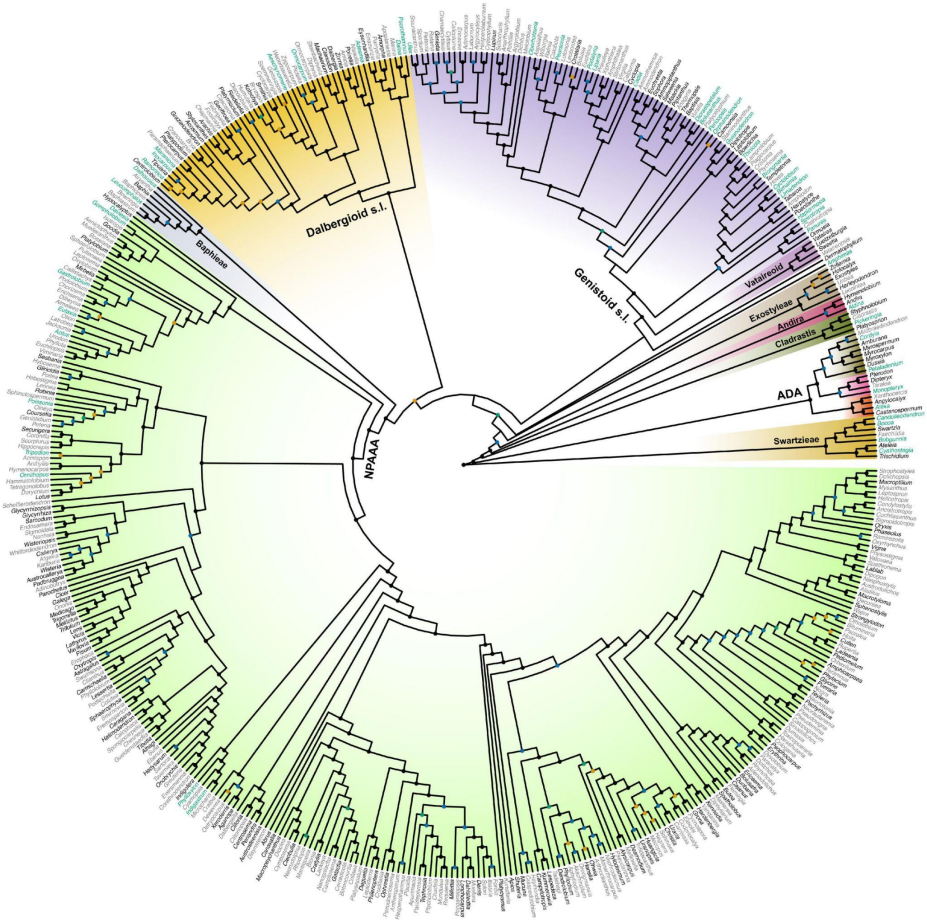


Figure 10 The recent Papilionoideae phylogeny produced by maximum likelihood.

IRLC and Robinoids, in the sense used in this thesis, are within the green part of the tree. ADA: Angylocalyceae, Dipterygeae, and Amburaneae. NPAAA: non-protein amino acid accumulating (Choi et al., 2022).

D. Rhizobia

Rhizobia is an umbrella term that refers to a diverse polyphyletic group of Gram-negative bacteria found in the phylum of Proteobacteria and composed mainly of alpha- and beta-proteobacteria (Willems, 2006) live in the soil but are also able establish nitrogen-fixing symbiosis with legume plants **(Figure 11)**. Historically, rhizobia were thought to be limited to a few genera in the class of alpha-proteobacteria (Checcucci et al., 2019). However, recent

studies have shown that rhizobia also include beta-proteobacteria, such as *Burkholderia* and *Cupriavidus,* that are able to establish an effective symbiosis with certain legumes and form root nodules (Gehlot et al., 2013).

Rhizobia belonging to the class of alpha-proteobacteria are members of the genera *Sinorhizobium (Ensifer), Bradyrhizobium, Azorhizobium, Microvirga, Neorhizobium, Phyllobacterium, Pararhizobium, Devosia, Methylobacterium, Aminobacter, Shinella, Mesorhizobium* and *Ochrobactrum.* The rhizobia among beta-proteobacteria include the genera *Burkholderia, Ralstonia,* and *Cupriavidus* **(Figure 11)**. Research about rhizobia has gained extensive interest due to their important role in agriculture as an essential provider of nitrogen to legume plants (Poole et al., 2018).

The ability of rhizobia to fix atmospheric nitrogen and establish a symbiosis with legumes is a complex process involving a multistep signal exchange process governed by a set of symbiotic genes **(Figure 11)**. This process starts first by detecting the presence of the host using specific genes, invasion of the host root and nodule development and then fixing atmospheric nitrogen using the nitrogenase enzyme. The genes responsible for these different steps are found clustered in the bacterial chromosome associated with bacterial transposons (Arashida et al., 2022) or in plasmids. These symbiotic genes are spread and transferred between rhizobia by both vertical and horizontal gene transfer (Z. Liu et al., 2019).

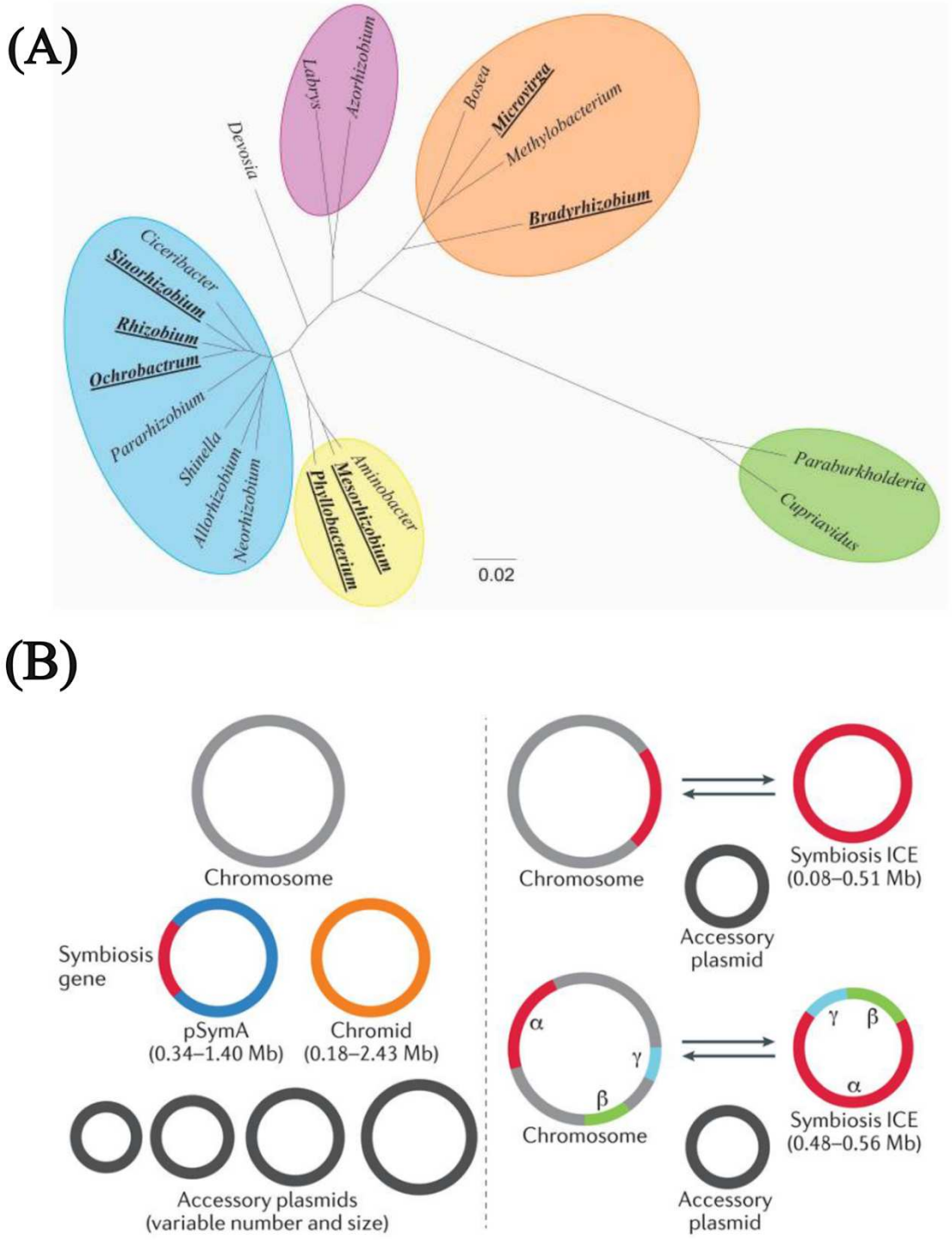


Figure 11 The phylogenetic distribution of rhizobia and the genomic organization of the symbiotic genes in the rhizobia genome.

(A) A consensus tree of rhizobia based on 16S rRNA from (Stępkowski et al., 2018), where the polyphyletic group of rhizobia includes species from beta-proteobacteria (the green clade) and alpha-proteobacteria (the rest). (B) The genomic organization of the symbiotic genes in rhizobia from (Poole et al., 2018), where they can be present in the plasmids like in *Sinorhizobium meliloti* or chromosomal (in symbiotic islands) like in *Bradyrhizobium* or *Mesorhizobium*.

E. The molecular dialogue of the symbiotic infection between legume plants and rhizobia

The symbiotic interaction between legumes and rhizobia is carried out through different processes, namely bacterial infection, nodule formation, nitrogen fixation, and bacteroid differentiation. First, the rhizobial infection is initiated by the exchange of specific signaling molecules between legumes and rhizobia. When legume plants are in nitrogen starvation, they produce and release flavonoids, a group of phenolic secondary metabolites, in the rhizosphere. These molecules attract rhizobia, initiating the formation of nodules (Heidstra et al., 1994). The presence of compatible flavonoids in the rhizosphere the symbiotic signaling cascade by activating of nodulation (nod) genes. This activation leads to the synthesis and secretion of lipochitooligosaccharides called Nod factors (Mergaert et al., 1997) that are recognized by specific membrane receptors in legume plants, called LysM-receptor-like kinases, such as LYK3 and NFP in *Medicago* and NFR1 and NFR5 in *Lotus*, which form heterodimer complexes (Heckmann et al., 2006). Legume receptors directly bind to rhizobial Nod factors, inducing the formation of root nodules that will house the nitrogen-fixing bacteroids.

After the molecular dialog between legume and rhizobia is established, the bacterial infection and nodule organogenesis initiate, where a few bacteria are trapped into the plant tissue by the curling root hair formed by the legume in response to this molecular dialog. These few bacteria constitute the founding cells that will become a complete nodule population. The rhizobia will then penetrate the root hair cell through the infection thread resulting from the curling root hair **(Figure 12)**. The infection thread works as a conduit for rhizobia to travel from the first penetration to the incipient nodule meristem (Brewin, 1991). When the infection thread reaches a young nodule cell the rhizobia is released into the plant cell by endocytosis (Newcomb & Wood, 1986) to form symbiosomes (Skorupska et al., 2006). These symbiosomes are organelle-like structures surrounded by a plant-derived membrane, in which rhizobia grow, divide, and differentiate into nitrogen-fixing bacteroids.

When rhizobia infect a plant cell, it undergoes a shift from dividing to differentiating into a nitrogen-fixing cell, leading to polyploidy. This polyploidy is due to multiple cycles of endoreduplication, where genome replication occurs without cell division (Mergaert et al., 2006). This phenomenon leads to the enlargement of these plant cells, which leads to the extreme sizes observed in the fixing zone (Mergaert et al., 2006).

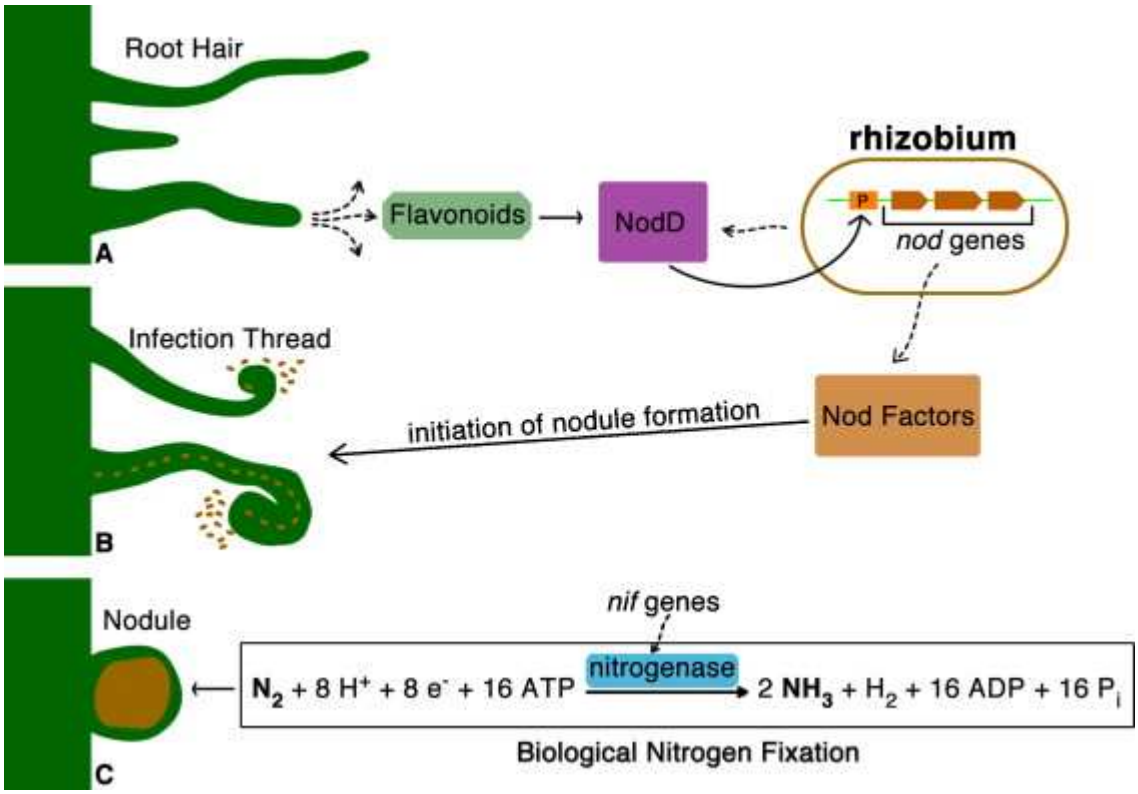


Figure 12 Overview of the nodulation and biological nitrogen fixation processes in the symbiosis between legume and rhizobia.

(A) The bacterial infection initiates the symbiosis process where the legume plant secretes flavonoids that trigger the rhizobia's synthesis and secretion of Nod factors. (B) After the recognition of Nod factors by Nod factor receptors, the nodule formation is initiated by the formation of an infection thread where rhizobia progress until the formation of the mature nodule. (C) The bacteria are housed inside the nodule in symbiosomes where they convert atmospheric nitrogen into (Laranjo et al., 2014).

The nitrogen-fixing mature nodules contain the uninfected peripheral nodule tissues and the nitrogen-fixing central tissue (**Figure 13**). The central tissue contains plant cells infected with rhizobia and uninfected cells. The peripheral tissues are the endodermis, cortex, and parenchyma. There are two different types of nodules, determinate and indeterminate, that differ in development and structure (Hirsch, 1992). Indeterminate nodules, such as in alfalfa, clover, and pea, are initiated from the inner cortex and form a persistent nodule meristem at their apex that allows for continuous growth and cell division, resulting in elongated nodules with distinct zonation (**Figure 13**) (Xiao et al., 2014). In contrast, determinate nodules, such as in soybean and bean, develop from the outer cortical cells and lack a persistent meristem that is present only in the early stages of development. This results in spherical nodules with limited meristematic activity. In addition to the difference in the meristem, other features distinguish

between the two structures. The indeterminate nodule has a specific zonation with different parts, from the meristem to the root (**Figure 13**).

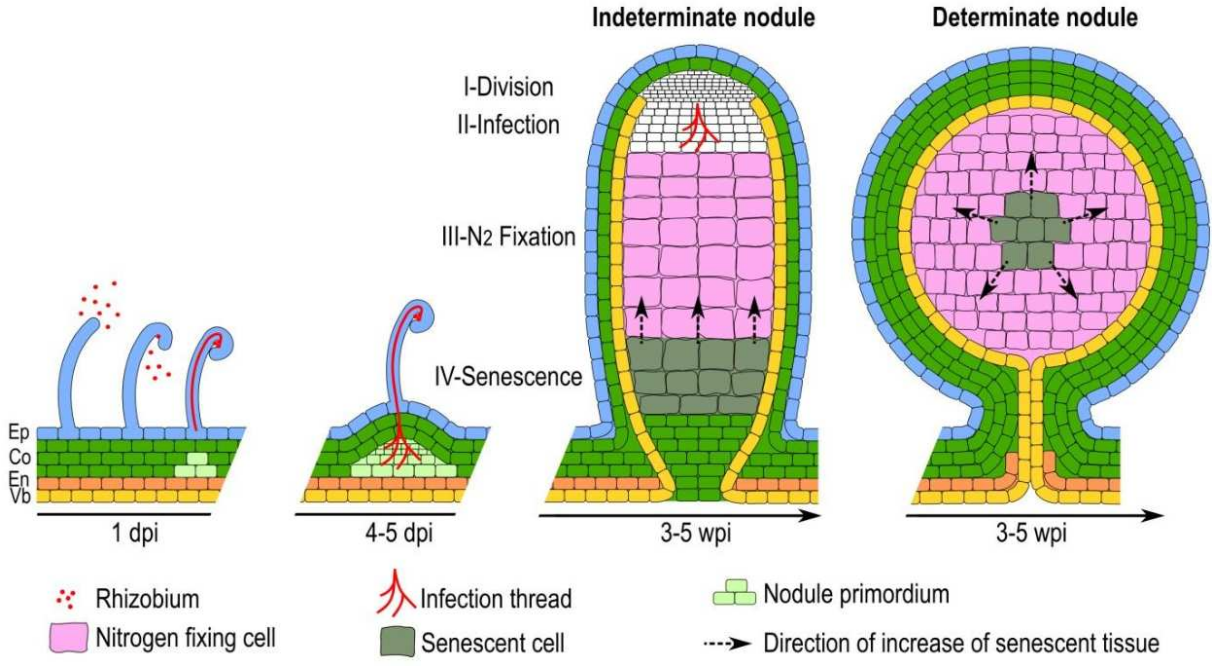


Figure 13 The differences between determinate and indeterminate nodules.

The legume nodules can be indeterminate with an elongated shape and different zones: division or meristem (zone I), infection (zone II), nitrogen-fixing (zone III), and senescence (zone IV), or determinate with a spherical shape (Kazmierczak et al., 2020). The white cells in the indeterminate nodules are meristematic. En, endodermis; Co, cortex; dpi, day post-inoculation; Ep, epidermis; Vb, vascular bundle; wpi, week post-inoculation.

The enzymatic complex responsible for the process of nitrogen fixation is the nitrogenase. In the mature nodules, inside the bacteroids, the genes encoding nitrogen fixation and respiratory functions are activated and differentially expressed in response to low oxygen concentration to initiate the nitrogen fixation process (Roux et al., 2014). The nitrogenase complex is known for its sensitivity to dioxygen, where high concentrations of dioxygen in the nodule irreversibly inactivate this complex. However, rhizobia need oxygen for cellular respiration to produce energy. Thus, it is crucial to maintain low levels of dioxygen and prevent the inactivation of the nitrogenase. This role is achieved by the leghemoglobin, which acts as an oxygen buffer. The absence of leghemoglobin can lead to early nodule senescence and failure in nitrogen fixation (L. Wang et al., 2019).

3. Terminal Bacteroid Differentiation (TBD) within nodules

In some legume plants, after the formation of mature nodules, the nitrogen-fixing bacteroids inside them undergo an extreme bacterial differentiation called Terminal Bacteroid Differentiation (TBD). Terminally differentiated bacteroids fix nitrogen more efficiently than undifferentiated ones, generating greater gains in plant biomass and reproduction (Oono & Denison, 2010).

A. The different bacteroid features that change during TBD

In legume-rhizobium symbiosis, the rhizobia inside the nodule are housed inside the symbiosome, a membrane-bound compartment of the legume host. Inside the symbiosome these bacteria become nitrogen-fixing bacteroids. Interestingly, the shape, physiology, and nitrogen fixation efficiency are different from one symbiotic program to another. In some symbiotic programs, the bacteroids resemble free-living (culture) bacteria with unaltered shape and morphology (**Figure 14**). This type of bacteroid is called U-type for unmodified or non-swollen (**Figure 14**). The bacteroids of *Bradyrhizobium diazoefficiens* USDA110 in the nodules of the *Glycine max* exemplify the U-shape (**Figure 14**). However, in some legume plants bacteroids undergo an extreme differentiation program called Terminal Bacteroid Differentiation (TBD), becoming elongated (E-shape) or spherical (S-shape), also called swollen bacteroids (**Figure 14)**. Terminally differentiated bacteroids irreversibly lose the ability to divide and this process appears to be controlled by the host plant (Mergaert et al., 2006; Van de Velde et al., 2010). TBD is characteristic of five legume clades: IRLC, Dalbergioids, Genistoids, Indigoferoids and Mirbelioids, suggesting that the ability to cause TBD has arisen several times independently in the legumes (Oono & Denison, 2010). To date there are no reports of TBD in the Robinoid or Millettioid clades (Oono & Denison, 2010). Indeed, some rhizobia strains can form terminally differentiated bacteroids or not, depending on the host legume plant. For example, the IRLC legumes generally host E-shape bacteroids, such as *Medicago truncatula* (**Figure 14**) and *Medicago sativa* (Vasse et al., 1990). However, there are few IRLC species that host S-shape bacteroids, such as *Ononis spinosa* (Montiel et al., 2016). In Dalbergioids clade, we can find either plants that host spherical or elongated shapes. For example, *Arachis hypogaea* and *Aeschynomene indica* host spherical bacteroids (**Figure 14**). Yet, no studies have been conducted to analyze this TBD in detail at cellular and molecular levels in the other TBD-inducing clades.

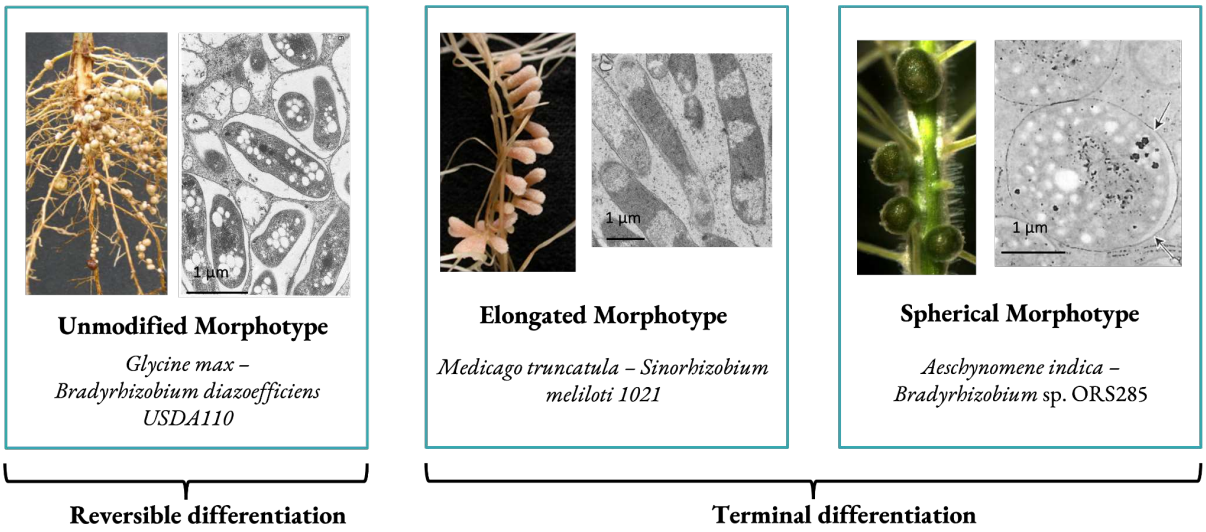


Figure 14 The diversity in the morphology and the shape of bacteroids during bacteroid differentiation in legume-rhizobia symbiosis.

In some symbiotic programs, the bacteroids retain their original morphology (left box). In other symbiotic programs bacteroids become elongated or spherical and polyploid (the two right boxes).

Terminally differentiated bacteroids are larger, have a permeabilized membrane (**Figure 15A**), and are polyploid through endoreduplication without cell division (**Figure 15**) (Haag et al., 2011). This process is governed by the *ctrA* gene (**Figure 16**). CtrA is the master regulator of the bacterial cell cycle, which controls different cellular processes, such as chromosome replication and cell division (Pini et al., 2015). Moreover, a novel cell cycle regulator called *fcrX* has been identified that directly acts on both CtrA and FtsZ, thereby controlling cell cycle, division, and symbiotic Terminal Bacteroid Differentiation (Dendene et al., 2022). FcrX is required for the establishment of symbiosis in *Medicago truncatula.* Overexpressing *fcrX* increases the symbiosis efficiency and induces the TBD state earlier (Dendene et al., 2023).

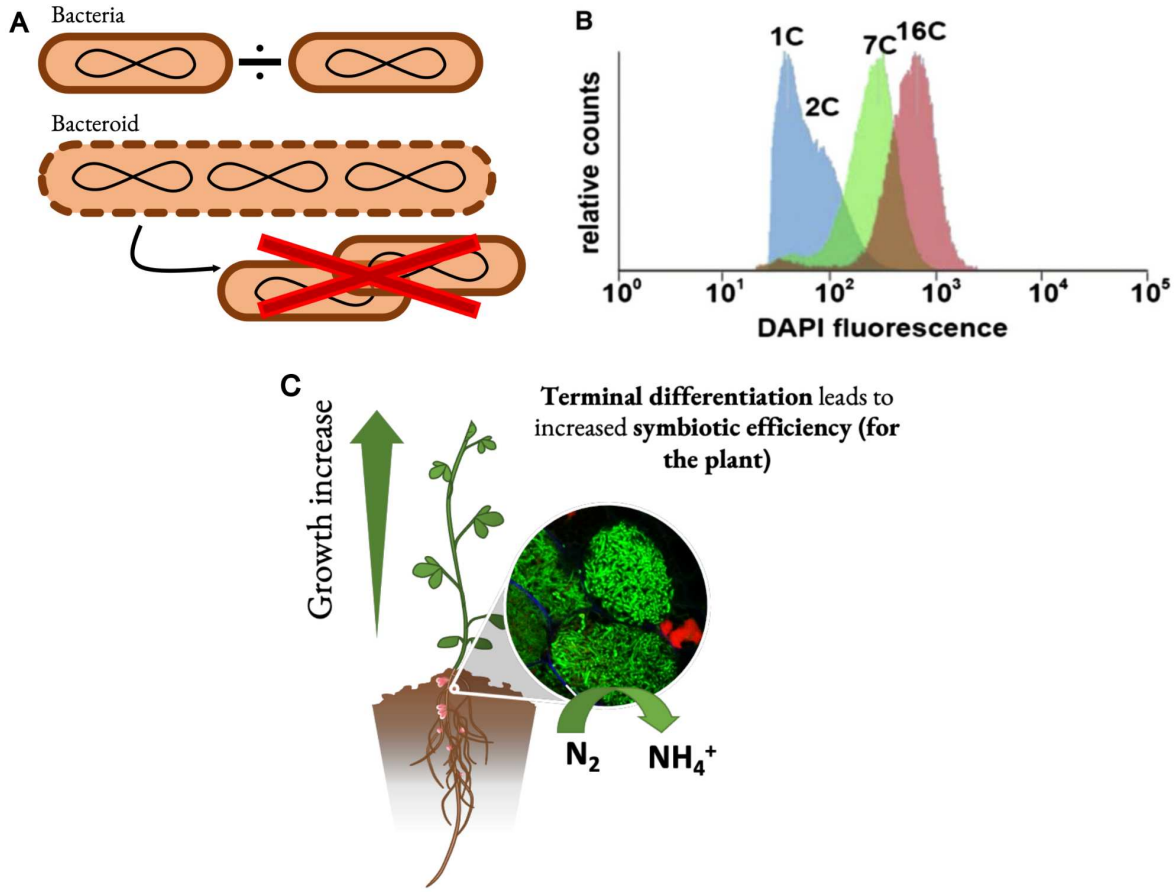


Figure 15 TBD enhances symbiotic efficiency.

(A) During TBD, the bacteroids undergo endoreduplication and do not divide again, resulting in increased cell volume and an elongated or spherical shape. (B) The DNA content of free-living *Bradyrhizobium* ORS285 bacteria (blue) and bacteroids from different host plants (*Aeschynomene afraspera* in green and *Aeschynomene indica* in red), measured by flow cytometry after DAPI staining from (Czernic et al., 2015). (C) The TBD increases nitrogen fixation efficiency, resulting in greener leaves and increased plant growth (adapted from (Nicoud et al., 2021)).

B. TBD increases the symbiotic efficiency

Various studies have shown that differentiated bacteroids are more efficient at nitrogen fixation, which increases the plant biomass (**Figure 15**), than non-differentiated ones (Oono & Denison, 2010). These studies compared two host plants with the same rhizobial strain, one inducing TBD and the other not. For example, when pea and bean legumes are inoculated with the same rhizobial strain, the pea plants that induce TBD have increased nitrogen fixation supported by increased plant biomass compared to the bean plants that do not induce TBD (Oono & Denison, 2010). Other studies have extended these findings by demonstrating that there is a positive correlation between the extent of TBD and the symbiosis efficiency (Kazmierczak et al., 2017). Indeed, when *Medicago* hosts are inoculated with different rhizobial strains, they show a

positive correlation between the terminal bacteroid differentiation (endoreduplication extent and cell enlargement in the bacteroids) and the symbiotic efficiency (Kazmierczak et al., 2017). Moreover, a study on *Aeschynomene* species from the Dalbergioids clade has shown that the spherical bacteroids that undergo a strong differentiation are more efficient in symbiosis compared to less differentiated elongated ones (Lamouche, Gully, et al., 2019).

C. NCR peptides induce TBD

The Terminal Bacteroid Differentiation process involves various modifications, such as membrane permeabilization, inhibition of bacterial division, DNA amplification, and cell enlargement, all orchestrated by the action of plant peptides called NCR (Nodule-specific Cysteine-Rich) peptides (Mergaert, 2018). NCR peptides are identified to induce TBD based on transcriptomic and genomic comparisons of legume species (Mergaert et al., 2003). These peptides were first identified in IRLC legumes, where *Medicago truncatula* has the best-studied NCR family with more than 600 different NCR peptides (Montiel et al., 2017). NCR peptides are secreted by the host plant and targeted to the bacteroids, indicating a peptide-based mechanism for bacteroid differentiation under host control (**Figure 16B**) (Van de Velde et al., 2010). They are specifically (differentially) expressed in nodules (Guefrachi et al., 2014). Further, it has been shown that NCR genes are expressed in different zones of the nodule (**Figure 16A**) and thus are involved in different stages of TBD (Guefrachi et al., 2014).

NCR peptides are small peptides composed of a signal peptide that allows their secretion and a mature peptide with 30-60 amino acids. Besides the 4, 6 or 8 conserved cysteines (**Figure 16B**) in the mature peptide, the amino acid composition of NCR peptides is highly diverse (Czernic et al., 2015). This diversity in amino acid sequences of NCR peptides results in a wide variety of peptides with different properties, including antimicrobial activity.

NCR peptides are related to antimicrobial peptides called defensins, which are part of the innate immune response (Guefrachi et al., 2014; Horváth et al., 2015). The mature peptides of plant defensins are 40-75 amino acids long, and they also have a diverse amino acid composition despite the 8 conserved cysteines (**Figure 16B**).

NCR peptides are present in all studied IRLC species (*Medicago sativa, Medicago truncatula, Pisum sativum, Cicer arietinum, Astragalus canadensis, Glycyrrhiza uralensis, Oxytropis lamberti, Onobrychis viciifolia, Galega orientalis,* and *Ononis spinosa*). However, their number is variable, from a few peptides in *Glycyrrhiza uralensis* to 700 in *Medicago truncatula*

(Montiel et al., 2017). All these NCRs from the IRLC clade have a conserved cysteine motif with four or six conserved cysteines (**Figure 16B**). An interesting fact is that the efficiency of TBD is positively correlated to the number of NCR peptides (Montiel et al., 2017). Moreover, the legume plants that do not induce TBD, such as *Lotus japonicus* and *Glycine max,* do not produce NCR peptides (Mergaert et al., 2003). Another study has also highlighted the presence of NCR-like peptides in the Dalbergioids clade where TBD occurs (Czernic et al., 2015). These NCRs have been identified from the nodule transcriptome of *Aeschynomene* species, which are different from the IRLC NCRs but exhibit similar functions in TBD (Czernic et al., 2015). The number of *Aeschynomene* NCR peptides ranges from 40 to 80 peptides, depending on the species (Czernic et al., 2015; Guefrachi et al., 2015). However, the amino acid sequences of these NCR peptides are distinct from the IRLC NCR sequences and are highly diverse. Indeed, two different motifs were identified in *Aeschynomene* NCR protein sequences. The first motif has a defensin signature with 8 conserved cysteines (type-2 NCR), while the second motif resembles the IRLC NCR motif with 6 conserved cysteines (type-1 NCR) (Czernic et al., 2015). As in the IRLC species, the transcriptomics analysis showed that *Aeschynomene* NCR genes are differentially expressed in the nodules, and the proteomics analysis revealed that NCR proteins are present in the bacteroids (Haag et al., 2011).

The involvement of NCR peptides in TBD has been validated with experimental assays, where blocking the transport of NCR peptides to the bacteroids leads to the failure of TBD (Maróti & Kondorosi, 2014). In contrast, the expression of NCR genes in species where TBD does not occur, such as *Lotus japonicus,* leads to bacteroids that mimic TBD with similar features (Van de Velde et al., 2010).

According to the isoelectric point, NCR peptides can be cationic, anionic, or neutral. The cationic NCR peptides disrupt the bacterial lipid membrane integrity and increase cell permeability (Tiricz et al., 2013; Van de Velde et al., 2010). These peptides, like other antimicrobial peptides, interact with negatively charged bacterial membranes, affecting membrane integrity (Lima et al., 2022). The disruption caused by NCR peptides can lead to bacterial death by inducing membrane damage and permeabilization (Haag et al., 2011). Interestingly, during TBD, NCR peptides do not kill the bacterial symbionts. Instead, they reach their intracellular targets without inducing cell death. For example, the treatment of *Sinorhizobium meliloti* (symbiotic rhizobium) with NCR247 *in vitro* leads to massive transcriptome alterations, affecting a substantial portion of bacterial life, including critical cell cycle regulators and cell division genes (Penterman et al., 2014). For cell cycle regulation,

different studies suggest that NCR peptides are implicated in interfering with the cell cycle regulatory network, leading to a decrease in CtrA expression during Terminal Bacteroid Differentiation (**Figure 16C**) (Lamouche, Bonadé-Bottino, et al., 2019; Roy et al., 2020). Indeed, the treatment of free-living bacteria with NCR247 is sufficient to decrease the expression of *ctrA* and other regulators (Penterman et al., 2014). Furthermore, NCR peptides have been suggested to provoke a cell cycle switch in symbiosis by targeting CtrA (Dendene et al., 2022). For cell division, it has been shown that NCR247 can inhibit bacterial cell division by interacting with the FtsZ protein (**Figure 16C**) (Farkas et al., 2014).

Moreover, it has been shown that NCR247 interacts with other bacterial proteins, such as ribosomal proteins and the chaperonin GroEL (**Figure 16C**) (Farkas et al., 2014). Studies have shown that exposure to sublethal doses of NCR247 leads to the downregulation of genes related to ribosomal subunits, suggesting a role in ribosome diversification (Farkas et al., 2014). Additionally, NCR247 treatment activates the expression of *rpoH1* and *rpoH2*-regulated genes (**Figure 16C**), mimicking the effects of lethal doses of NCR247 (Tiricz et al., 2013).

Despite the demonstrated importance of cationic NCR peptides, most NCR peptides found in the bacteroids are anionic and neutral, suggesting that they also have an important role in TBD (Durgo et al., 2015). For example, the NCR-like peptides found in *Aeschynomene* species are almost neutral and anionic, and none of them display antimicrobial activity *in vitro* against *Bradyrhizobium sp*. ORS285 (Czernic et al., 2015).

To induce TBD, NCR peptides are expressed by the host plant in the nodules and targeted to the symbiosomes where the bacteroids are housed. They are synthesized and translocated across the endoplasmic reticulum (ER), where a signal peptidase cleaves the signal peptide to release the mature peptide (Maróti et al., 2015). NCR transcripts are translated by the ribosomes in the ER, where the NCR peptides are synthesized and folded in their 3D structure. After the cleavage of the signal peptide by the signal peptidase complex, they are transferred to the Golgi apparatus, where they are embedded into secretory vesicles. These vesicles are transported to the symbiosome and fused with them, where the NCR peptides are then released in the bacteroids (**Figure 16C**) (Alunni & Gourion, 2016; Van de Velde et al., 2010). The importance of NCR peptides in TBD has been validated again by the defect in nitrogen fixation caused by the NCR secretory pathway *dnf* mutants. In *Medicago truncatula*, three *dnf* mutants defective in signal peptidase complex have been identified (Horváth et al., 2015; Kim et al., 2015). These three mutants, including *dnf1, dnf4,* and *dnf7,* exhibit various deficiencies in TBD, leading to

impaired nitrogen fixation. However, the *dnf1* mutant is the most defective because this mutation affects the cleavage of the signal peptides. Therefore, even if the ER synthesizes the NCR peptides, they are not cleaved and, thus, not transported to the bacteroids (Van de Velde et al., 2010). The two other mutants, *dnf4* and *dnf7,* are responsible for mutations in the NCR genes encoding for NCR211 and NCR169, respectively (Horváth et al., 2015; Kim et al., 2015). These mutations have been shown to lead to premature senescence of nodules and the death of bacteroids, either before (Horváth et al., 2015) or after the establishment of TBD (Kim et al., 2015). Furthermore, the downregulation by RNA interference in the DNF1 homolog of *Aeschynomene evenia, AeDNF1,* has been shown to lead to a significant defect in Terminal Bacteroid Differentiation (Czernic et al., 2015).

Even though *Medicago truncatula* has various NCR peptides, it has been shown that individual NCR peptides are essential for the establishment of TBD. These essential NCRs include the above-mentioned NCR211 and NCR169 (Horváth et al., 2015; Kim et al., 2015), where the mutation in the genes encoding for these peptides displays a strong defective phenotype. NCR247 is a pivotal NCR peptide that plays a central role in the establishment and maintenance of TBD. NCR247 has been shown to trigger TBD of *Sinorhizobium meliloti* both *in vitro* and *in planta* (Van de Velde et al., 2010). While NCR169, NCR211, and NCR247 were initially identified as critical peptides for TBD, recent research has highlighted the importance of additional peptides, NCR343 and NCR-new35 (Horváth et al., 2023). These studies demonstrated that anionic NCR peptides are also essential for TBD, where NCR211, NCR343, and NCR-new35 are anionic (Horváth et al., 2023). Furthermore, it has been demonstrated that NCR peptides are expressed in waves during different stages of nodule formation and bacteroid differentiation (Guefrachi et al., 2014).

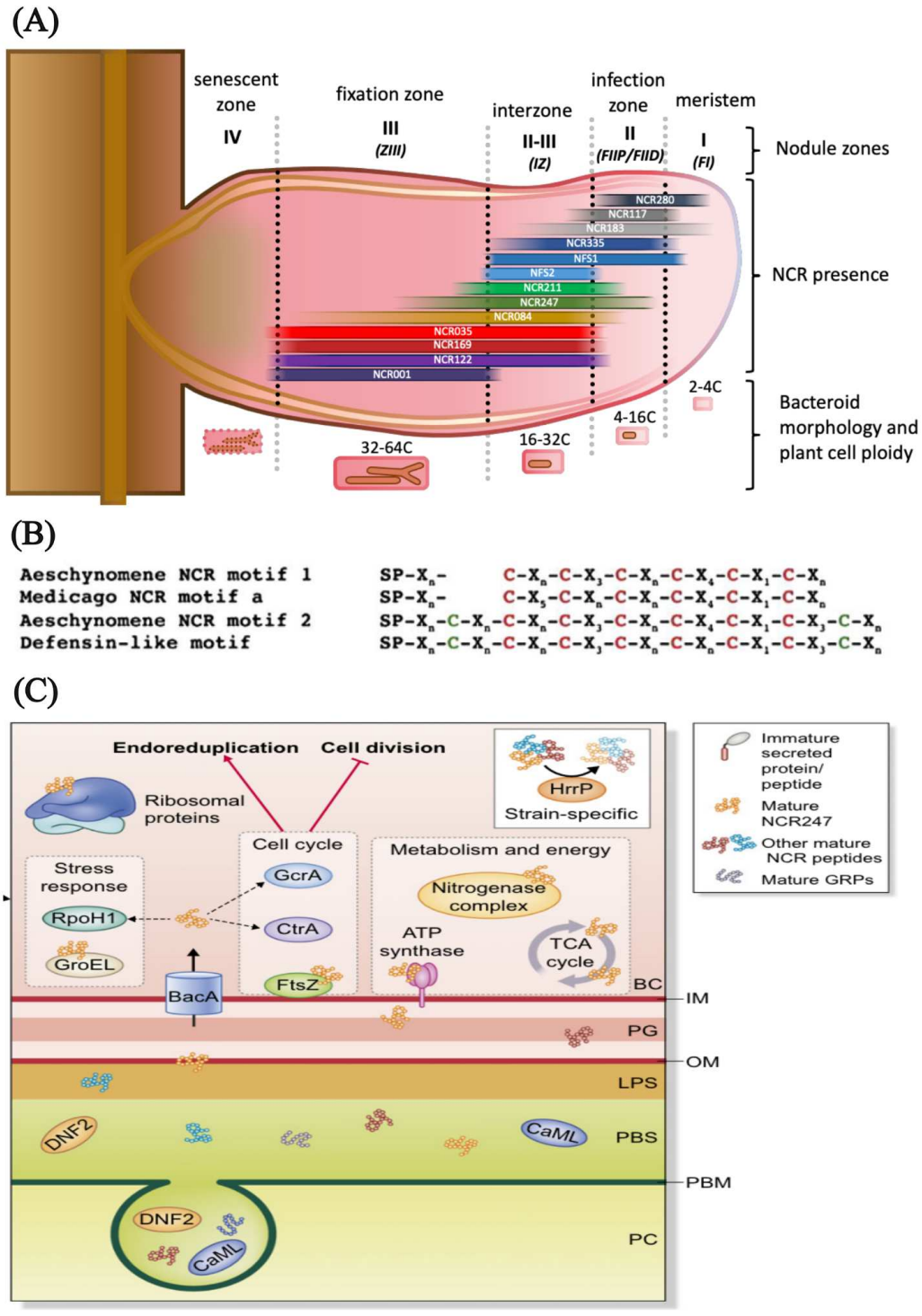


Figure 16 NCRs are the orchestrators of Terminal Bacteroid Differentiation, where they regulate various processes.

(A) Schematic representation of the elongated nodule of *Medicago truncatula* legume in symbiosis with *Sinorhizobium meliloti* leads to TBD induced by NCR peptides expressed in the different zones of the nodule. (B) The backbone structure and the cysteine motif of NCR peptides form IRLC (*Medicago*) and Dalbergioids (*Aeschynomene*) clades, and Defensin peptides (Czernic et al., 2015). (C) NCR peptides are secreted from the host legume plant and targeted to the bacteroid that became elongated through TBD (here in the case of *Medicago truncatula* and *Sinorhizobium meliloti*) from (Alunni & Gourion, 2016).

Furthermore, it has been demonstrated recently that NCR peptides are involved in iron binding, where they form complexes with heam, facilitating iron uptake by rhizobia (Sankari et al., 2022).

Recently, NCR peptides have been identified in different other IRLC and Dalbergioids species. Transcriptomic analysis of *Arachis hypogaea* (peanut) from the Dalbergioids clade identified 55 NCR peptides in this species (Raul et al., 2021). However, all identified NCRs are motif 2 (defensin-like) with 8 conserved cysteines in the mature peptide. The absence of motif 1 NCRs, despite their presence in *Aeschynomene* species, highlights the diversification of NCR amino acid sequences, even in the same clade (Raul et al., 2021). In the IRLC clade, a recent transcriptomic analysis of *Melilotus officinalis* (clover) predicted a new set of 308 NCR peptides in this legume species (Huang et al., 2022). Interestingly, among them, only 40 NCRs have sequence similarities with *Medicago* NCRs, and none of them have similarities to the two important *Medicago truncatula* NCRs (NCR211 and NCR169) (Huang et al., 2022), which highlights again the rapid diversification of NCR peptides. Another recent study (Dinkins et al., 2022) identified 425 NCR peptides in *Trifolium pratense* (red clover), an IRLC legume. Notably, this species has the highest percentage of cationic NCR peptides among all studied IRLC species, with 32% of the NCR peptides being cationic (Dinkins et al., 2022). Furthermore, based on transcriptomic analysis, 167 NCR peptides were reported in the *Astragalus sinicus* IRLC legume (Wei et al., 2022). This research found one important *Astragalus sinicus* NCR that displays bacterial growth inhibition *in vitro* and defective symbiotic phenotype *in planta* (Wei et al., 2022). They also identified an interaction between GroEL proteins and two *Astragalus sinicus* NCR peptides, which is consistent with the previous finding demonstrating that *Medicago truncatula* NCR247 has a direct interaction with GroEL proteins (Farkas et al., 2014). Moreover, AHL transcription factor orthologs have been found to be essential for NCR gene expression and regulation in *Medicago truncatula*, which opens a way to improve nitrogen fixation efficiency in legumes lacking NCRs (S. Zhang et al., 2023).

In summary, NCR peptides are pivotal effectors that trigger the process of Terminal Bacteroid Differentiation in the symbiotic interaction between legumes and rhizobia. Their role in governing the TBD is essential for the establishment of functional nitrogen-fixing nodules. The complex process of TBD involves the coordinated action of NCR peptides, impacting various cellular pathways within bacteroids (**Figure 16C**). However, NCR peptides in other clades where TBD occurs remain unexplored.

D. BacA transporters are essential for TBD

In order to protect the bacteroid membranes from the damaging antimicrobial peptide activity of cationic NCR peptides, bacterial ABC transporters are involved in the process of Terminal Bacteroid Differentiation. The most important transporter involved in this process is the BacA transporter, a membrane ABC transporter encoded by the *bacA* gene. This transporter is a homolog of *Escherichia coli* SbmA transporter, and BacA-like transporters (BclA) have also been identified in other organisms. The SbmA_BacA or BclA transporters belong to the ABC transporters family. BacA transporter was identified and suggested to be involved in the transport of peptides across the membrane more than thirty years ago, while NCR peptides were still unknown (Glazebrook et al., 1993). This study reported the first *bacA* mutant that failed to fix nitrogen in legumes, highlighting the critical role of the BacA protein in nitrogen fixation during symbiosis with legumes (Glazebrook et al., 1993). The ABC transporters are known to have three sub-units: the Periplasmic-binding Protein (PBP), two transmembrane domains, and one ATPase domain. In contrast to the classic ABC transporters and BclA transporters, BacA lacks the ATPase domain and has a different mechanism of import compared to the ATP-hydrolyzing ABC transporters (Travin et al., 2022). BacA protein is a transmembrane protein with seven transmembrane domains that import molecules inside the inner membrane using the proton-motive force (LeVier et al., 2000).

BacA and BclA proteins play important roles in diverse biological processes and, more precisely, in transport mechanisms. This family of transporters is involved in the import of different molecules, such as antimicrobial peptides and antibiotics (Slotboom et al., 2020). These transporters are essential for surviving antimicrobial peptides like Bac7, where they play an important role in importing them (Arnold et al., 2013). They are also implicated in the transport of antibiotics inside the bacterial cells (Ferguson et al., 2002).

BacA transporter is also known to be involved in symbiotic interactions, such as in *Sinorhizobium meliloti* and *Brucella abortus*, where it affects lipid-A fatty acids (Ferguson et al., 2004). They are involved particularly in bacterial interactions with eukaryotic hosts, where it has been shown that BacA is involved in the uptake of eukaryotic peptides in *Sinorhizobium meliloti* (Marlow et al., 2009). Furthermore, it has been shown that BacA homologous proteins have an important role in importing molecules involved in host-pathogen interactions, which validates their importance in host defense (Arnold et al., 2013). For instance, studies have shown that BacA is crucial for the long-term survival and persistence of various pathogens,

including *Brucella abortus* and *Mycobacterium tuberculosis* (Arnold et al., 2013; Wehmeier et al., 2010). In *M. tuberculosis*, BacA transporters have been implicated in the maintenance of chronic infections in murine models (Arnold et al., 2013; Domenech et al., 2009; Marlow et al., 2009). It has been suggested that BacA may function as an importer without requiring a substrate-binding protein (SBP) or that the SBP might be expressed elsewhere in the genome and interact with BacA to form a complete ABC transporter (Haag et al., 2013). Moreover, BacA has been identified as the sole transporter for cobalamin and corrinoids (Gopinath et al., 2013). In *B. abortus,* studies have shown that BacA transporters are essential for the establishment of chronic intracellular infections within mammalian hosts (Marlow et al., 2009). The BacA protein affects the very long-chain fatty acid (VLCFA) modification of lipopolysaccharides (LPS) in both *S. meliloti* and *B. abortus*, which is vital for the chronic intracellular infections underlying their pathogenesis (Ferguson et al., 2004). Additionally, the BacA protein is involved in the uptake of peptides. It is necessary for the establishment of chronic intracellular infections by *S. meliloti* and *B. abortus* within their respective hosts (Wehmeier et al., 2010).

Moreover, the Sbma BacA in *Escherichia coli* is known for the uptake of antimicrobial peptides (Travin et al., 2022), where it has been shown that the *sbmA* mutants exhibit different phenotypes of resistance and sensitivity of glycine-rich antimicrobial peptides, Bac7, and bleomycin (Mattiuzzo et al., 2007). This transporter is also involved in the uptake of microcin C, peptide-nucleotide antibiotics (Nicoud et al., 2021). In addition to that, the BacA transporter is involved in importing drugs, which causes DNA damage and cell death (LeVier & Walker, 2001).

However, it has been shown that *bacA* mutants exhibit increased sensitivity to diverse membrane stresses, such as ethanol, antibiotics, and detergents (Arnold et al., 2013; Ferguson et al., 2002, 2004; Haag et al., 2011; Ichige & Walker, 1997; Nicoud et al., 2021), which suggests that the antimicrobial peptide transport activity of BacA transporter is not the unique source of the *bacA* mutant phenotypes. These mutants are characterized by defects in membrane integrity, altered cell envelope structure, and reduced resistance to different stresses found in the host environment, suggesting that the BacA transporter is essential for maintaining the integrity of the bacterial cell envelope (Arnold et al., 2013; Ferguson et al., 2002, 2004). In addition, the altered cell envelope of *bacA* mutants makes them more susceptible to environmental stresses, such as membrane-damaging agents and exposure to acidic pH (Bellaire et al., 2005; Domenech et al., 2009; Haag et al., 2011). Furthermore, it has been shown that

bacA mutants are linked to reductions in outer membrane lipid content and, more precisely, in very long-chain fatty acids (VLCFAs), which have an impact on membrane stability and resistance to stresses (Marlow et al., 2009). It has been shown that the role of the BacA transporter in maintaining membrane integrity is also essential for survival in symbiotic interactions of rhizobia bacteria, such as *Sinorhizobium meliloti* (Haag et al., 2011). In summary, the increased sensitivity of *bacA* mutants to diverse stresses highlights the importance of BacA in protecting the bacteria from environmental challenges and in promoting their adaptation to different host environments (Ferguson et al., 2004; Haag et al., 2011; Karunakaran et al., 2010).

During legume-rhizobia nitrogen-fixing symbiosis, BacA and BclA transporters are essential for bacteroid differentiation induced by NCR peptides, leading to the Terminal Bacteroid Differentiation process in the bacteroids. It has been shown that BacA transporter from *Sinorhizobium meliloti* is required for bacteroid differentiation in symbiosis with *Medicago* spp. (Alunni & Gourion, 2016) and its homologous gene *bclA* in *Bradyrhizobium* is also essential for TBD in *Aeschynomene* spp. (Czernic et al., 2015). While rhizobial *bacA* mutants exhibit similar phenotypes in free-living conditions, it has been shown that *bacA* mutants are unable to fix nitrogen when in symbiosis with IRLC legumes, such as *Medicago truncatula* and *Medicago sativa* **(Figure 17)** (Glazebrook et al., 1993; Haag et al., 2011; Nicoud et al., 2021). NCR peptides rapidly kill them upon release from the infection threads **(Figure 17)**.

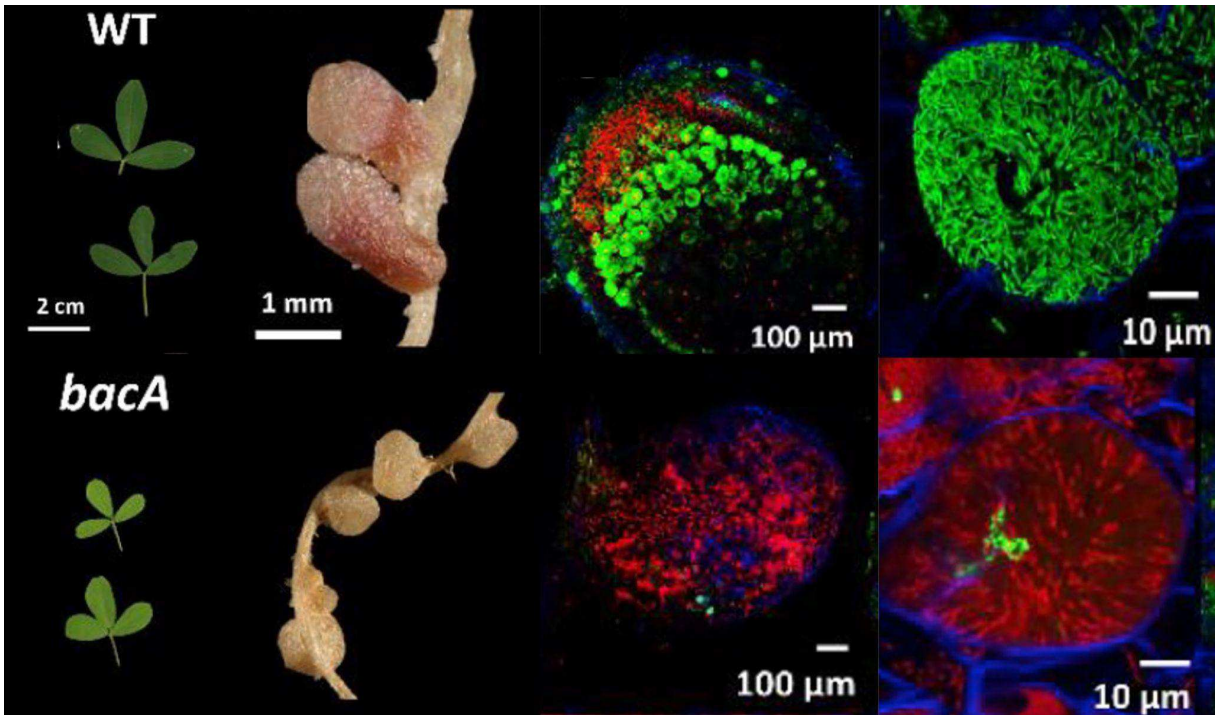


Figure 17 Highly reduced survivability of Sinorhizobium meliloti bacA mutants.

The WT and *bacA* mutant phenotypes of *S. meliloti in planta* in symbiosis with *Medicago sativa.* The leaves are greener and bigger, the nodules are bigger and pinker in the WT condition, and the live cells are present almost only in the WT condition. Adapted from (Nicoud et al., 2021).

Interestingly, BacA transporter is essential only in TBD in symbiosis with legume species that produce NCR peptides, where *bacA* mutants survive within legume hosts where TBD does not occur, such as *Lotus japonicus* and *Phaseolus vulgaris* (Haag et al., 2013) **(Figure 18).**

According to these studies, it has been suggested that the requirement of BacA in NCR-dependent symbiosis that induces TBD may be to avoid membrane damage by driving away the NCR peptides from the bacterial membrane (Nicoud et al., 2021) and to import NCR peptides to their intracellular targets to induce TBD (diCenzo et al., 2017; Haag et al., 2011). Indeed, it has been shown that BacA is required to survive NCR peptides *in vitro* (Haag et al., 2011).

BclA is a homolog of BacA transporter in *Bradyrhizobium* species, which is also required for TBD in NCR-triggered symbiosis with Dalbergioids species such as *Aeschynomene* spp. (Guefrachi et al., 2015). This transporter has a different structure and transport mechanism from BacA, where it possesses an ATPase domain that allows import by ATP hydrolysis.

The *bclA* mutants in the *Bradyrhizobium* strains ORS285 and USDA110 (Barrière et al., 2017) in symbiosis with legumes that do not (soybean) or do (Aeschynomene) impose TBD revealed the requirement of BclA transporter only in NCR-producing plants. However, the function of BacA and BclA *in planta* may differ, where, first, in *bclA* mutants, the bacteria are not killed immediately upon release (Guefrachi et al., 2015). Second, while the *Sinorhizobium meliloti bacA* did not complement the *bclA* mutant of *Bradyrhizobium* strain ORS285, the *bclA* gene of ORS285 partially complements the defect of the *S. meliloti bacA* mutant (Guefrachi et al., 2015). These results agree with the difference in the structure and the transport mechanisms of these two proteins.

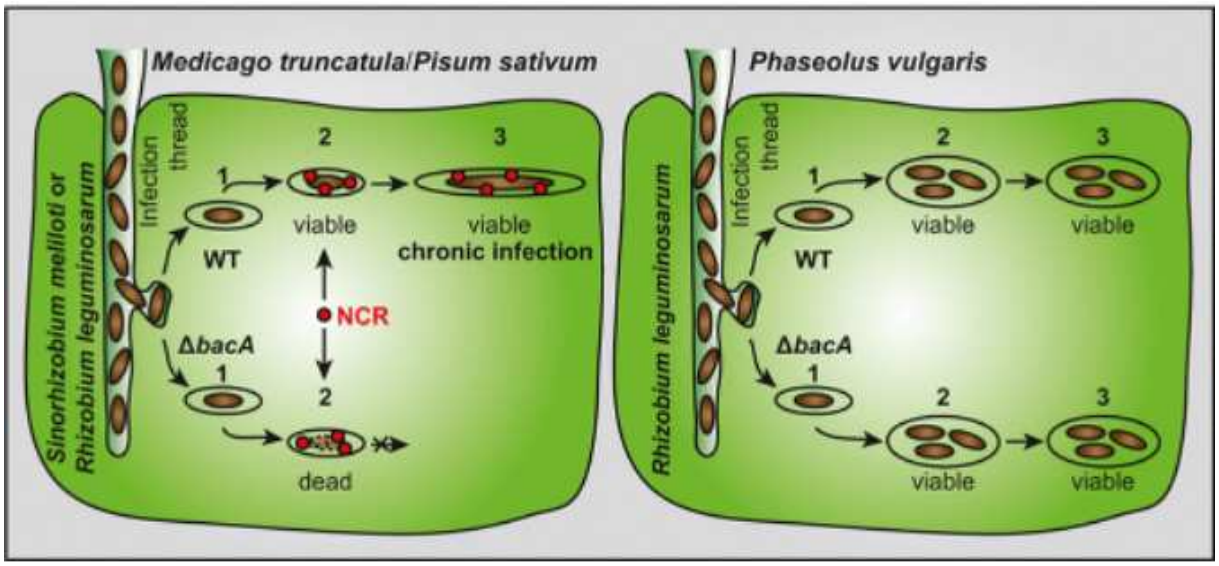


Figure 18 BacA transporter is required only in legume-rhizobia symbioses that involve the TBD process.

The sensitivity of bacterial *bacA* mutants against NCR peptides, where we see that a BacA transporter is required for chronic infection and survival of antimicrobial peptides (Haag et al., 2011).

BacA transporters possess homologous in several bacteria, including plant symbionts and human pathogens. This conservation of BacA homologs may indicate that they are not specific and are involved in peptide transport for diverse functions in bacteria (Glazebrook et al., 1993). It has also shown that the defect caused by the *sbmA* mutant in *Escherichia coli* can be complemented by *Sinorhizobium meliloti* BacA (Domenech et al., 2009).

The importance of these transporters relies on their capacity to transport NCR peptides, which is crucial for bacterial survival inside the host environment because NCR peptides, as said before, are antimicrobial peptides that cause damage to the bacterial membrane (Barrière et al., 2017).

In accordance with the requirement of BacA transporter in NCR peptide-induced TBD, we know that all rhizobial symbionts of IRLC and Dalbergioids encode BacA or BclA. Despite their structural and import mechanism differences **(Figure 19)**, the BacA and BclA similarities extend to their functions in symbiotic relationships with hosts. According to the work of (Guefrachi et al., 2015), we can see **(Figure 19)** that five homologous groups represent BacA. However, only two distant homologous of them are involved in symbiosis, BacA, and BclA. In addition to being phylogenetically distant, these two transporters have different structures and methods of transport. Yet, both proteins can import NCR peptides, allowing TBD-promoting nitrogen fixation. However, the differences in managing NCR peptides by BacA and BclA suggest that they may have evolved in response to specific host interactions, producing different NCR peptides (Barrière et al., 2017).

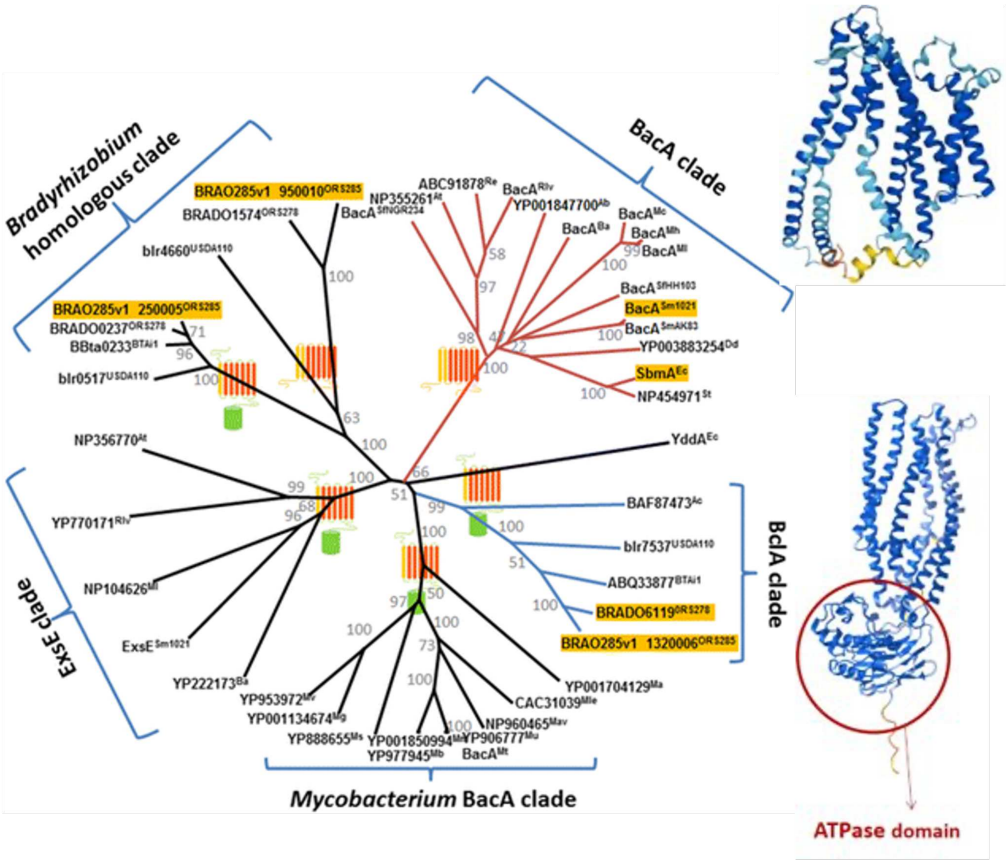


Figure 19 BacA and BclA transporters are phylogenetically and structurally distinct, but they are involved in the same symbiotic program of TBD.

Phylogenetic tree of BacA homologous proteins. Blue and red branches represent the transporters involved in symbiosis. The green box represents the proteins that possess the ATPase domain. The red square represents the proteins with the transmembrane domain (from Guefrachi et al. 2015). The structures on the right represent the Alphafold2 structures of BacA (top) without the ATPase domain and BclA (bottom) with the ATPase domain (Guefrachi et al., 2015).

E. YejABEF protein - another transporter of NCR peptides

The YejABEF is another ABC transporter involved in Terminal Bacteroid Differentiation in legume-rhizobia symbiosis. YejABEF protein has been identified as an important inner membrane ABC transporter conserved in Gram-negative bacteria, involved in antimicrobial peptide resistance (Couturier et al., 2022). Different studies have shown that YejABEF is involved in the import of different antimicrobial peptides, such as microcin C, which contributes to the resistance of bacteria against these peptides (Eswarappa et al., 2008; Vondenhoff et al., 2011). The importance of the YejABEF transporter in bacterial survival was highlighted by demonstrating its requirement for the virulence of pathogens, such as *Brucella melitensis* (Z. Wang et al., 2016). Additionally, the *yejABEF* mutant in *Salmonella* is more sensitive to peptides that provoke membrane damage, such as defensins and polymyxin B (Eswarappa et al., 2008).

Recent research (Nicoud et al., 2021) identified that YejABEF is essential for TBD in NCR-mediated symbiosis. The flow cytometry measurement of the DNA content of rhizobia isolated from nodules of the wild-type strain, *bacA* mutant, *yejA,* and *yejE* mutants shows that the TBD is affected in the Yej mutants **(Figure 20A).** However, the *yejABEF* cannot complement the function of the BacA transporter, where *bacA* mutants die directly upon release, but *yej* mutants display unusual phenotypes at the end of TBD (Nicoud et al., 2021). For instance, the *yejA, yejE,* and *yejF* mutants formed functional nodules (Nicoud et al., 2021). Yet, *yejA* and *yejF* mutants showed reduced nitrogen fixation activity (Nicoud et al., 2021). Additionally, while these three mutants displayed decreased cell viability, their nodules contained many differentiated and large bacteroids (Nicoud et al., 2021). This may be due to the different subsets of NCR imported by the two transporters.

In accordance with this, *in vitro* experiments show that all *yej* mutants are sensitive to at least one NCR peptide **(Figure 20B)**, highlighting the importance of this transporter in importing NCR peptides for an effective TBD (Nicoud et al., 2021). However, the *yejA* mutant displays a different *in vitro* sensitivity profile than the other mutants **(Figure 20B),** which is also associated with different symbiotic phenotypes (Nicoud et al., 2021), where it has similar nodules as the wildtype strain, and DNA content in the bacteroids near to the wildtype also **(Figure 20A)** (Nicoud et al., 2021).

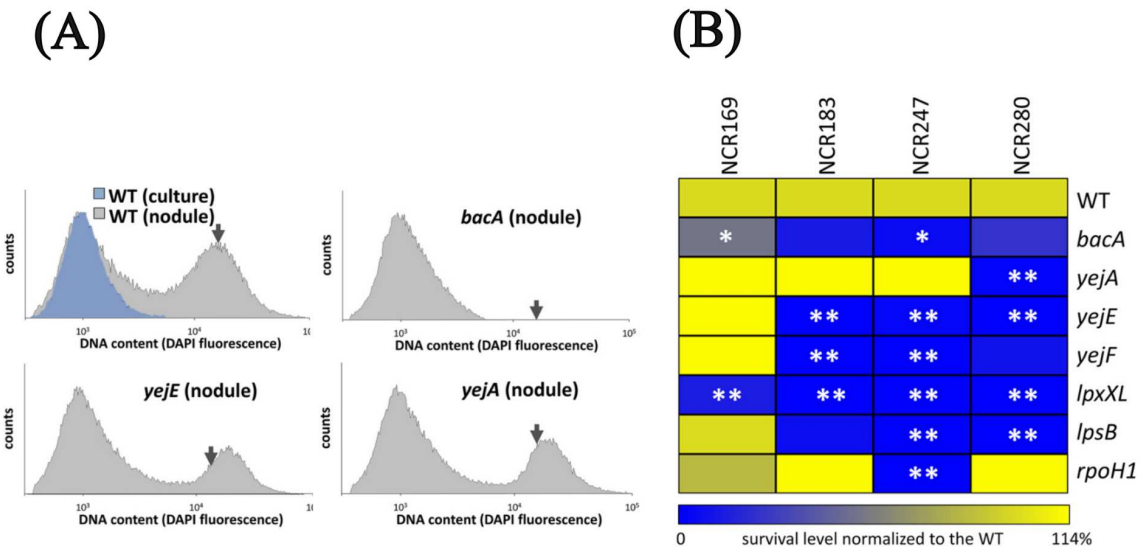


Figure 20 YejABEF transporter is essential for Terminal Bacteroid Differentiation

(A) Flow cytometry analysis that represents the DNA content of mutant strains isolated from *Medicago truncatula* nodules (B) Sensitivity of *Sinorhizobium meliloti* strains to different NCR peptides represented by the survival of mutant strains to the NCR treatment. The asterisks represent the significance of differences (Nicoud et al., 2021).

4. Evolution of Terminal Bacteroid Differentiation

Previously, it was assumed that the TBD triggered by NCR peptides is a specific feature of IRLC species. However, twelve years later, it has been shown that this feature is also present in Dalbergioids species, where NCR peptides have been identified in *Aeschynomene* species.

NCR peptides from the IRLC and Dalbergioid clades have different sequences and cysteine motifs, but both induce TBD in the symbiont. The evolution of NCR peptides in the IRLC clade with only type-1 NCRs is proposed to have originated from defensin ancestors (Mergaert et al., 2003). In contrast, the Dalbergioid NCRs with both type-1 and type-2 NCRs are different from IRLC NCRs (Czernic et al., 2015). Indeed, it has been suggested that NCR peptides evolved independently in IRLC and Dalbergioid clades, supporting the idea of convergent evolution driving symbiont differentiation (Downie & Kondorosi, 2021). However, a recent phylogenetic study between defensins and NCR peptides demonstrated that they may share the same origin (Salgado et al., 2022). Yet, NCRs and defensins are small peptides and have diverse amino acid sequences, and these studies are sequence-based and limited to a small subset of NCR peptides. Therefore, it remains to be elucidated if IRLC and Dalbergioid NCRs evolved from the same or different gene families. It is possible that they have evolved several times within the Papilionoideae clade or evolved from an ancestral gene, and they are expressed only in some

clades.

The process of Terminal Bacteroid Differentiation occurs in five legume clades (IRLC, Dalbergioids, Mirbelioids, Indigoferoids, and Genistoids) (**Figure 21**) (Oono et al., 2010). Nevertheless, only in two of them, IRLC and Dalbergioids, this process is known to be induced by NCR peptides on the legume side and BacA or BclA transporters on the bacterial side. However, the presence of NCR peptides in other clades that induce TBD, and their evolution remains unknown. The involvement of BacA and BclA in TBD and their evolutionary relationship remain to be studied. Moreover, currently, there is no study about the presence and distribution of BacA-like transporters in Bacteria.

In summary, to gain insights into the evolution of TBD and the independent coevolution in legume-rhizobia symbiosis, it is important to decipher the evolutionary history of TBD molecular actors, NCR peptides, and BacA-like transporters.

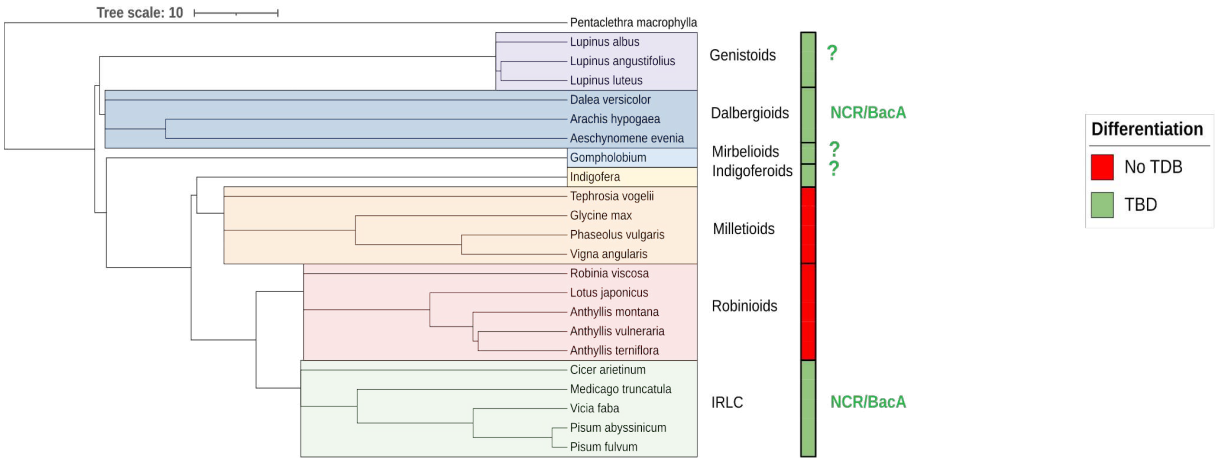


Figure 21 The presence of NCR peptides in other legume clades and the involvement of BacA and BclA in TBD remain unclear.

The phylogenetic tree of legume species created with TreeTime (Kumar et al., 2022) shows the distribution of Terminal Bacteroid Differentiation in legume-rhizobia symbiosis and the involvement of NCR peptides and BacA transporters in this process.

5. The approaches used to decipher the evolutionary history of NCR peptides and BacA transporters

Although it is common for the primary structure (amino acid sequence) to determine the biological function, similar shapes (tertiary and quaternary structures) and functions could be shared by proteins with highly diverse primary structures. We know that the function and the 3D structure of the proteins are more conserved than the amino acid sequences. In many instances, when we compare a protein family with the same biological function, we can see that even if the amino acid sequences are highly diverse, the 3D structures are similar.

A few studies have been conducted about the evolution of NCR peptides, and these studies used only sequence-based homology approaches. However, NCR peptides are small peptides (30-60 aa in the mature peptide) and highly diverse at the sequence level, even in the same clade (Czernic et al., 2015; Huang et al., 2022; Wei et al., 2022). On the other hand, no study has been conducted on the evolution of BacA and BclA transporters.

Therefore, in order to decipher the evolutionary histories of NCR peptides and BacA transporters, I implemented a bioinformatic pipeline based on statistical and structural bioinformatic tools, but also traditional homology-based and phylogenetic tools, such as Blast (Basic Local Alignment Search Tool) analysis and Maximum Likelihood tree inference.

A. Homology, orthology, and clustering

If two proteins share an evolutionary relationship, we say that these proteins are homologous. Among homologous proteins, we can distinguish two types: orthologs, which are homologs from different species, and paralogs, which are homologs from the same species that evolved from gene duplication. On the other hand, the commonality between the amino acid sequences represents the measure of similarity. Often, if two protein sequences have a considerable similarity, we say that they are homologous. A protein family is a group of proteins that share a common evolutionary origin and three-dimensional structures and often perform similar functions. These protein families are usually defined based on sequence similarity, structural comparison, and functional characteristics.

Thus, the first approach of the bioinformatic pipeline is the sequence-based homology (**Figure 22**). In this approach, the first method used is sequence alignment, which is an essential technique for finding regions of similarities between protein sequences. It involves arranging

sequences by inserting gaps to put regions of similarities in the same columns. There are various methods and algorithms for sequence alignment, such as pairwise sequence alignment, which includes the global alignment (Needleman-Wunsch algorithm) (Needleman & Wunsch, 1970) and the local alignment (Smith-Waterman algorithm) (T. F. Smith & Waterman, 1981); and Multiple Sequence Alignment (MSA). The tool used to score the sequence similarities here is the Blast software, which is an open-source computational tool that identifies regions of similarities between nucleotides and amino acid sequences using pairwise local alignment (Tatusova & Madden, 1999, p. 2). It is fundamental in identifying homologous sequences and is extensively used for sequence similarity searches. This tool can conduct a sequence similarity search for a sequence of interest (or multiple sequences) against a curated sequence database. The Blast software offers various approaches to perform sequence similarity analysis, such as Blastn, which searches a nucleotide database using a nucleotide sequence query, Blastp that searches an amino acid sequence query against a protein database, tBlastn which searches a protein sequence query against a translated nucleotide database and Blastx that searches a protein database with a translated nucleotide sequence query. Moreover, Blast is pivotal in all-versus-all searches to identify potential orthologs and protein families.

OrthAgogue is another sequence-based similarity tool that complements Blast by identifying putative orthologs and paralogs based on sequence similarity. OrthAgogue, a multithreaded C application, is a re-implementation of the second step of the OrthoMCL tool (L. Li et al., 2003), which is the identification of putative orthologs, aiming to improve the performance of orthology prediction in large datasets (Ekseth et al., 2014). This tool uses the all-versus-all Blast results and provides a similarity matrix that represents the orthologs and paralogs graph. Based on this similarity matrix, we can regroup the closely related proteins using a clustering algorithm, such as Markov Clustering (MCL) (Enright et al., 2002). The Markov Cluster Algorithm is a powerful clustering algorithm that has been widely used in bioinformatics and other fields to identify clusters in networks. MCL utilizes flow simulation to consider global relationships within a graph simultaneously during clustering, making it robust for separating diverged paralogs, distant orthologs, and sequences with different domain structures (L. Li et al., 2003). This approach based on sequence similarity using Blast, OrthAgogue, and MCL (**Figure 22**) allows us to regroup ortholog and paralog proteins from different species into no overlapping closely related protein clusters.

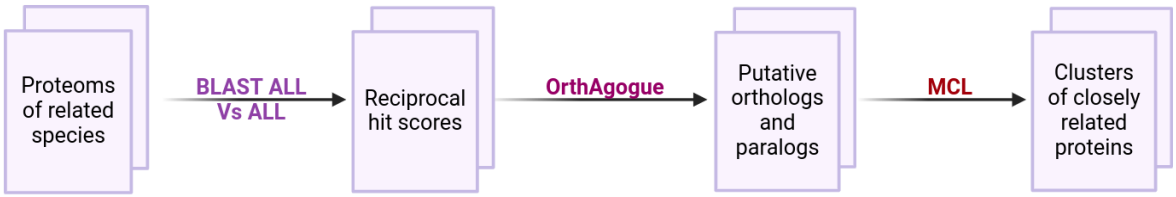


Figure 22 Workflow of the homology-based approach to identify and regroup closely related proteins.

This approach is similar to the OrthoMCL software (L. Li et al., 2003). The only difference is that orthAgogue is used instead of step two of OrthoMCL because of the time and memory consumption of this step.

B. Prediction and annotation based on Hidden Markov Model (statistical analysis)

The identification and classification of protein families are essential for understanding the relationships between proteins and predicting their functions, which provides insights into their evolutionary history and their functional diversity. Various methods are used to characterize protein families based on sequence alignments, Hidden Markov Models (HMM), and structural classifications. These methods play an important role in classifying proteins, annotating newly discovered proteins, and predicting their functions based on their relationships to known protein families.

Multiple sequence alignment (MSA) is a powerful tool that allows for the simultaneous alignment of three or more biological sequences, providing a more comprehensive analysis compared to pairwise sequence alignment. The MSA reveals the patterns of sequence conservation and the regions that are subjected to insertions or deletions. Constructing an MSA aids in identifying conserved motifs and protein domains that have structural and functional significance. The MSA is a fundamental technique in bioinformatics that is used in different tasks, such as phylogeny reconstruction, secondary structure prediction, and coevolution signal identification. A Hidden Markov Model (HMM) is a probabilistic model widely used in bioinformatics. In the context of MSA, the HMMs represent the distribution of probability of the sequences of observations. The HMM profile (S. R. Eddy, 1998) is the most used and powerful tool of HMM for searching databases efficiently. The HMM profiles are known for their ability to capture the presence or absence of motifs in sequences, making them a valuable tool for identifying remote homologs and assigning homologous sequences to curated protein families (Mistry et al., 2013). HMM profiles are more advanced than numerical profiles, where they can model insertions and deletions in a position-specific manner, providing a more qualified representation of sequence evolution (S. R. Eddy, 1998). The HMM profiles allow for

position-dependent gap penalties, and thus model better the variability in gap lengths in different positions in a sequence alignment, which can enhance the sensitivity and accuracy of sequence alignments. This feature is crucial in capturing the complex evolutionary relationships between sequences, especially in cases where traditional alignment methods may not be relevant.

Sequence similarity searches are essential for assigning new sequences to a protein family (Pearson, 2013). The classical methods, such as BLAST, identify sequences from databases that match a query sequence using local pairwise alignment. These searches can be improved by using Hidden Markov Models profile (pHMMs) to find new members of a protein family of interest by screening a sequence database using the pHMM (which models better the gaps) constructed from the MSA (which represents better the insertions and deletions), by aligning and scoring the matches.

HMMER software (Finn et al., 2011) is a widely recognized tool for sequence analysis, particularly for its probabilistic models to detect remote homologous. HMMER uses Hidden Markov Models (HMMs) to build profiles from MSAs, enabling sensitive and accurate sequence searches. The HMMER modules generally used are *hmmbuild, hmmsearch,* and *jackhammer.* The *hmmbuild* command is used to build HMM profiles from multiple sequence alignments. The *hmmsearch* command is used to search a sequence database using HMM profiles for hits with specific E-values, and the *jackhammer* is used to search a protein database iteratively using a sequence query.

While PFAM (Bateman et al., 2004) is a widely used database in association with HMMER, curated databases are also used with this software. The Pfam database is a valuable resource that contains a vast collection of protein-domain family multiple sequence alignments and Hidden Markov Models (HMMs) (Finn et al., 2014). Pfam consists of two main components: Pfam-A and Pfam-B. Pfam-A is composed of manually curated, well-characterized protein domain families with high-quality seed alignments, maintained through manual checks and family-specific Hidden Markov Model profiles (pHMMs) (Sonnhammer et al., 1997). On the other hand, Pfam-B was initially an automatically generated supplement to Pfam-A, clustering sequence segments not included in Pfam-A to enhance coverage (Bateman et al., 1999; Finn et al., 2010, 2015). The distinction between Pfam-A and Pfam-B lies in the curation process, with Pfam-A being manually curated and of higher quality, while Pfam-B is automatically generated (Mistry et al., 2013). Furthermore, Pfam's use extends to structural bioinformatics, where Pfam

families are assigned to protein structures in the Protein Data Bank (PDB), facilitating protein domain identification and functional annotation (Q. Xu & Dunbrack, 2012).

HMM profiles are integrated into different tools through HMMER software, such as the SPADA pipeline. SPADA (Small Peptide Alignment Discovery Application) is a homology-based gene-finding program specifically optimized for detecting and annotating small peptides with one or two exons (P. Zhou et al., 2013). This approach involves using MSAs of homologous genes within a gene family and building HMM profiles from them, which will be used to search genomes for this gene family using *hmmsearch* (**Figure 23**). On the other hand, this method uses gene prediction tools to assign the best candidate gene to each hit found with *hmmsearch* from the translated genomic sequences (**Figure 23**).

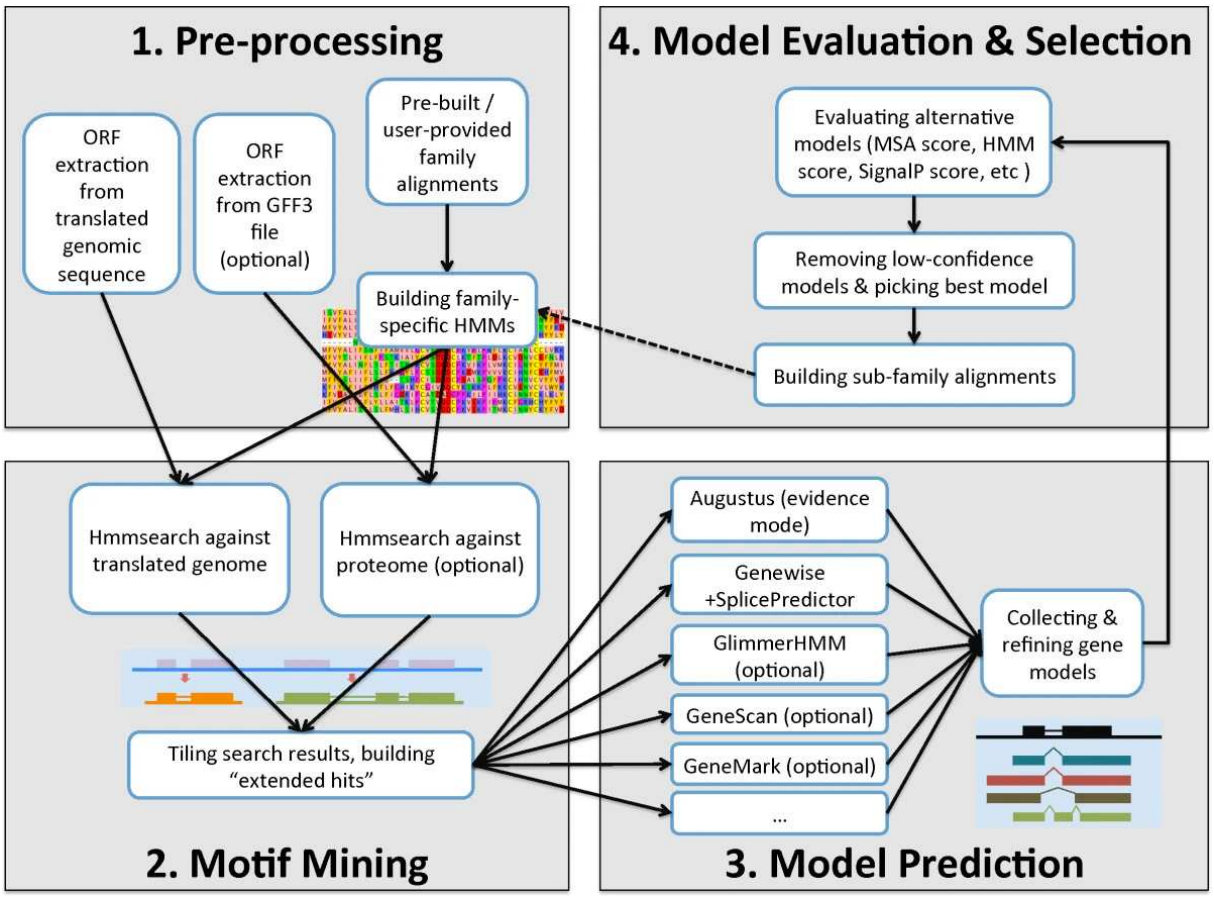


Figure 23 SPADA workflow.

(1) The first step of SPADA is to extract the ORFs from the genomes and to build HMM profiles of the protein family alignments. (2) The second step is to use hmmsearch to search for hits in the translated genome. (3) The third step is the evaluation step, where the hits are filtered according to MSA score, presence of signal peptide, etc. (4) The last step is to choose the best candidate gene for each hit using gene prediction tools such as Augustus and Genewiser and to remove the false positive using an E-value threshold. Figure from (P. Zhou et al., 2013).

In summary, combining profile HMMs and tools like HMMER, SPADA, and PFAM with classic sequence similarity approaches like BLAST can enhance the accuracy and sensitivity of identifying sequence family members. These methods are important for computational biology research, allowing researchers to reveal evolutionary relationships and gain insights into the function of the proteins from their sequences.

C. 3D structure prediction and structural phylogenetics

Proteins exhibiting similar functions may have conserved 3D structures despite sharing no overall sequence similarity. During evolution, it has been observed that protein structures are much more conserved than sequences (Bastolla et al., 2003). This conservation of protein structure is essential for maintaining the protein's function and can improve our understanding of protein evolution.

Protein structure prediction is a fundamental area in computational biology that focuses on determining the three-dimensional structure of a protein from its amino acid sequence. Predicting protein structure is challenging due to the computational complexity involved in determining the full 3D structure of a protein solely from its sequence. Over the years, significant progress has been made in this area, particularly with the growth of protein databases and advancements in computational methods, such as deep learning approaches like convolutional neural networks. While experimental methods like NMR spectroscopy and X-ray crystallography are traditionally used for accurate protein structure determination, computational methods have gained importance, particularly in monomer protein structure prediction, as evidenced by Critical Assessment of Structure Prediction (CASP) experiments (Moult et al., 2016).

CASP is the key platform for evaluating advancements in this field. CASP is a biennial worldwide competition that tasks participants with predicting the 3D structures of proteins based on their amino acid sequences (Kryshtafovych et al., 2019). *De novo* protein structure prediction methods have significantly advanced over the years, from the first CASP1 program initiated in 1994 (Moult et al., 1995) through CASP15. However, the folds were of poor quality, and the advancements were limited until CASP11, which focused on refining the template-based models (Modi & Dunbrack Jr., 2016). A notable advancement in CASP12 was the use of statistical methods that considered all pairs of residues simultaneously to address transitivity effects, leading to a substantial increase in accuracy, with an overall precision of 47%

(Kryshtafovych et al., 2019). In addition, there was an enhancement in predicting three-dimensional contacts between pairs of residues, which contributed significantly to the progress observed in CASP12 (Moult et al., 2018). Coevolution-based features and machine learning integration significantly boosted the average precision in CASP12, particularly in predicting long-range contacts (Adhikari et al., 2017). The improvements in CASP13 were multiple, including advancements in molecular dynamics simulations, deep learning methodologies, utilization of sparse data, contact prediction, and model refinement (Kryshtafovych et al., 2019). In CASP14, significant advancements were observed in protein structure prediction learning techniques, such as incorporating deep learning-based protein inter-residue distance predictors, has notably enhanced template-free tertiary structure prediction (J. Liu et al., 2022) **(Figure 24)**. The notable success in the CASP14 was AlphaFold2, an end-to-end deep learning method developed by DeepMind (Jumper et al., 2021) that has revolutionized protein structure prediction, achieving unprecedented modeling accuracy. The advancements in CASP14 have not only showcased the remarkable progress in protein structure prediction but have also highlighted the potential of deep learning, artificial intelligence, and novel algorithmic approaches to revolutionize the field and significantly enhance our understanding of protein structure. The CASP experiments not only focus on predicting protein structures but also extend to assessing ligand binding site predictions and modeling quaternary structures of protein-protein complexes (Kryshtafovych et al., 2021). Recently, they also predicted protein-DNA and protein-RNA complexes with CASP15 (Abramson et al., 2024; Kryshtafovych et al., 2023).

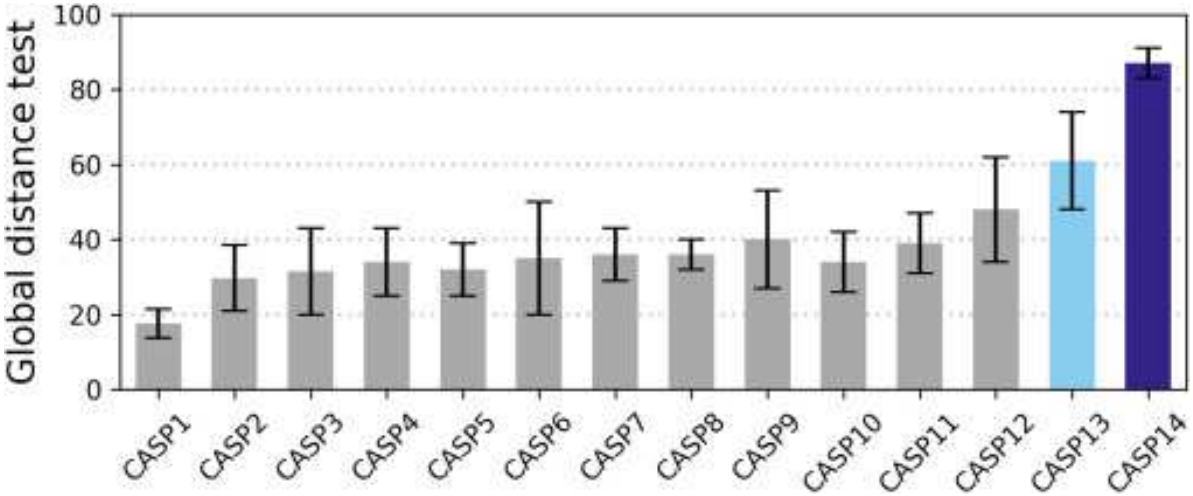


Figure 24 The global distance test (GDT) across the different CASPs.

The improvements in the global distance accuracy test. Figure from (Erdős & Dosztányi, 2023)

Developed by DeepMind, Alphafold, the most successful method of CASP14, has revolutionized structural biology by providing highly accurate protein structure predictions (Jumper et al., 2021). Alphafold utilizes machine learning and multiple sequence alignments to infer and refine pairwise residue-residue evolutionary and geometric information, achieving ground-breaking success in protein structure prediction (Jumper et al., 2021). This system employs deep neural networks to directly generate protein structure models from MSAs, achieving unprecedented accuracy. The latest version, Alphafold2, has notably outperformed its predecessor and other existing methods, demonstrating exceptional performance in the CASP14 competition (Jumper et al., 2021). The Alphafold2 program algorithms are trained with large protein structure databases such as the Protein Data Bank (PDB) and incorporate other sources of information (e.g., genetic sequences) to improve their predictions. This version employs an end-to-end deep neural network that takes MSA as input and outputs a structure model at the end. Alphafold2 uses information from amino acid sequences combined with MSAs and templates (similar structures) to predict protein structure. Alphafold2 highly depends on the number of MSAs and templates to provide accurate predictions. Indeed, large and high-quality MSAs provide more confident predictions.

Unlike its predecessor Alphafold, which used convolutional neural networks for distance map prediction, AlphaFold2 employs an end-to-end network where model parameters are jointly tuned from sequence input to structure output, optimizing the final model directly (Lupas et al., 2021). The crucial element of Alphafold2 is the Evoformer, which uses different pairwise modules to work on pairwise relations (distance, contact,...) between residues within the protein (Jumper et al., 2021) (**Figure 25**). Alphafold2 integrates neural network architecture and training procedures based on evolutionary, physical, and geometrical constraints of protein structures (Jumper et al., 2021). The model processes multiple sequence alignments (MSA) and templates through a translation and rotation equivariant transformer architecture, producing 3D structural models (**Figure 25**).

The learning algorithm's capacity to perform accurate protein structure predictions is measured using the Global Distance Test (GDT) score ranging from 0 to 100, which compares the predicted structure to the actual structure of a known protein. A GDT score of 100 indicates a perfect match, while a score of 0 indicates no match between the predicted and experimentally determined structures. In the CASP14 experiment, Alphafold2 achieved an average GDT score of 92.4 for all targets, with some predictions achieving scores as high as 99.9. This level of accuracy was previously thought to be achievable only through experimental methods (e.g., X-

ray crystallography or cryo-electron microscopy).

Furthermore, Alphafold provides confidence metrics like the predicted Local Distance Difference Test (pLDDT) ranging from 0 to 100 to assess how well the predictions align with experimental structures, enhancing result reliability (Mariani et al., 2013). Regions with a pLDDT score above 90 (> 90) indicate that the region has been modeled with high confidence and accuracy. Regions with pLDDT scores between 70 and 90 are considered generally confident (good) predictions, while regions with pLDDT scores between 50 and 70 have low confidence and should be interpreted with caution. Structural data from regions with pLDDT scores below 50 (< 50) are predictors of disorder. They may be unstructured under physiological conditions or may achieve structure when included as part of a complex (Varadi et al., 2022).

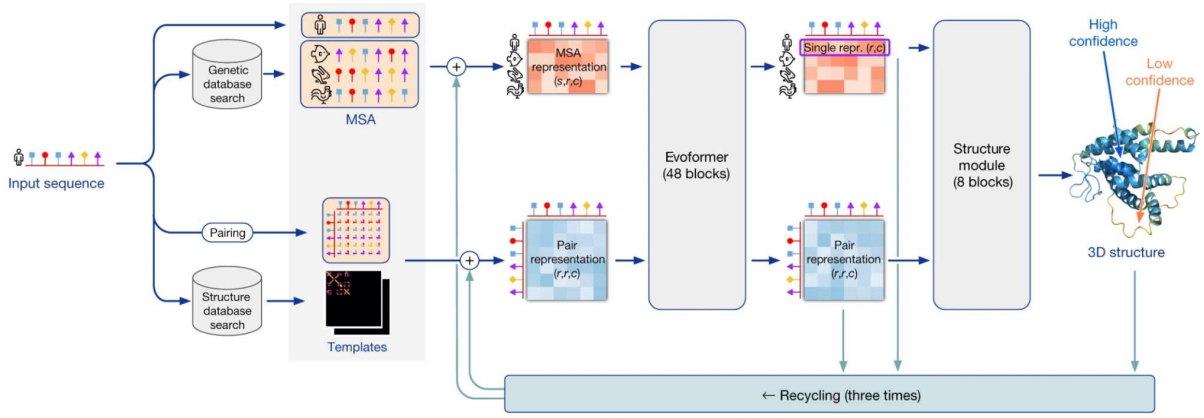


Figure 25 The workflow (architecture) of Alphafold2.

The arrows represent the information flow among the various components. Array shapes are shown in parentheses with s, the number of sequences; r, the number of residues; and c, the number of channels. Figure from (Jumper et al., 2021).

Foldseek is a novel tool that facilitates the rapid comparison of extensive structural data sets with a high degree of accuracy and sensitivity, aiding in the identification of structural similarities among different protein structures (van Kempen et al., 2024). Foldseek uses a structural alphabet to describe the tertiary interactions between amino acids (**Figure 26**). The 20 states included in the 3D interaction alphabet (3Di) describe the geometric conformation of each residue with respect to its spatially nearest neighbor. Specifically, Foldseek converts the query structures into 3Di sequences and then uses a pre-trained substitution matrix to search through the 3Di sequences of the target structures. This model allows Foldseek to be faster than other structural comparison tools, enabling large-scale protein structure comparison and clustering. For example, Foldseek achieves 86% of the sensitivity of DALI (Holm, 2022)

(Distance-Matrix Alignment), known to be the most sensitive structural alignment tool, and works 4,000 to 184,600 times faster (van Kempen et al., 2024).

Foldseek aligns the structure of a query protein against a database by representing tertiary amino acid interactions as sequences over a structural alphabet (van Kempen et al., 2024). By encoding protein structures in a sequence representation, Foldseek facilitates rapid alignment using MMseqs and local sequence alignment algorithms (**Figure 26**).

Foldseek has been applied in various ways, such as conducting structural alignment and clustering across databases like Alphafold DB (Barrio-Hernandez et al., 2023), identifying distant homologs, assigning functional annotations by integrating sequence-based and structure-based methods to identify structural orthologs and resolving conflicting predictions (Monzon et al., 2022), and clustering protein structures.

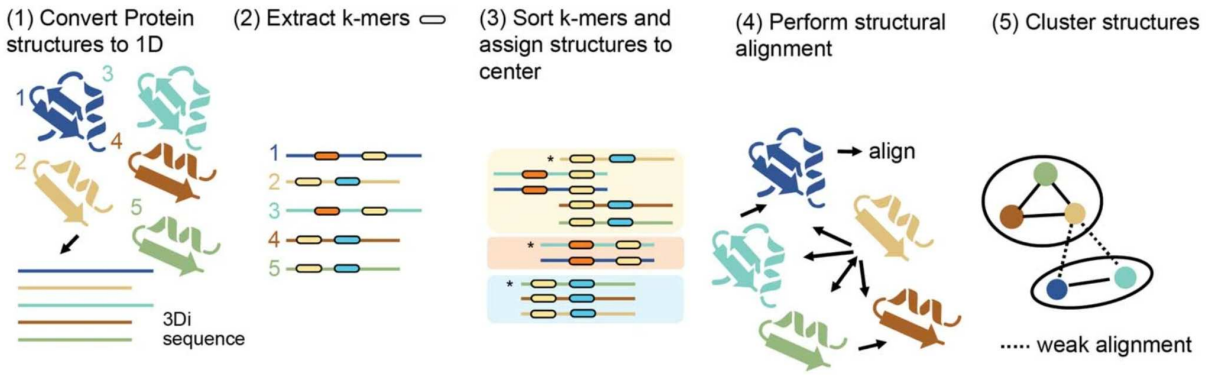


Figure 26 The five-step structural clustering approach using Foldseek's 3Di alphabet.

(1) Protein structures are converted to 3Di sequences and processed through the Linclust workflow. (2) For each sequence, 300 min-hashing k-mers are extracted and sorted. (3) The longest structure is assigned to be the center of each k-mer cluster. (4) Structural alignment is performed in two stages: first, an ungapped alignment based on shared diagonal information is performed, hits are pre-clustered, and second, the remaining sequences are aligned using Foldseek's structural Smith-Waterman. (5) The remaining structures meeting alignment criteria are clustered using MMseqs2's clustering module. After the Linclust step the centroids are successively clustered by three cascaded steps of prefiltering, structural Smith-Waterman alignment, and clustering using Foldseek's search. Figure from (Barrio-Hernandez et al., 2023).

Furthermore, Foldseek has also been utilized for clustering structures based on TM-score calculations. Because the TM (Template Modelling) system is designed to score larger distance errors as weaker than smaller ones, the values obtained are more sensitive to global fold similarity than local structural variations. A TM-score of 1.0 indicates a perfect match between two structures, while scores of ≤ 0.17 represent random (unrelated) protein pairs. Protein pairs

that have a TM-score > 0.5 are considered significantly similar and expected to exhibit the same folding (J. Xu & Zhang, 2010).

In summary, Foldseek's innovative approach of treating structural alignment as a sequence alignment problem has been commended for its efficiency and effectiveness in comparing protein structures. This strategy has been pivotal in aligning protein structures at scale, enabling the identification of commonalities and unique features in protein structure space across different organisms (Bordin et al., 2023).

Foldtree is a structural phylogenetic tool that uses Foldseek to perform an all-vs-all comparison of protein structures and build a structural distances matrix (Foldseek structural sequence identity is used) that is used to build a distance-based phylogenetic tree (**Figure 27**) (Moi et al., 2023).

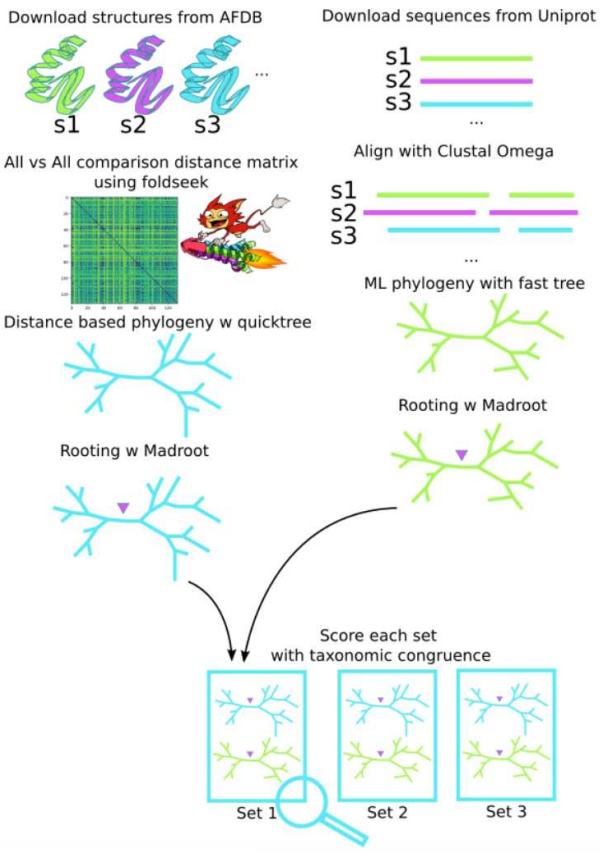


Figure 27 Foldtree schematic pipeline

Trees are created from equivalent protein sets for structure and sequence trees. On the structural side of the pipeline, all vs all comparisons using Foldseek are compiled into a distance matrix. These are used as input for Quick Tree and rooted with MAD. On the sequence side of the pipeline, the sequences are aligned with ClustalO or Muscle. A maximum likelihood tree is derived using fasttree or IQTree and then rooted with MAD. Figure from (Moi et al., 2023).

D. Phylogenetic analysis

Phylogenetics is the study that aims to understand the relationship between different species based on their evolutionary history and to reconstruct the diversification history from the ancestral population LUCA (Last Universal Common Ancestor) to actual organisms after billion years of population (Betts et al., 2018). It involves the study of molecular and morphological data to infer the evolutionary history of living organisms.

Phylogenetic trees are essential tools in evolutionary biology, providing a graphical representation of the evolutionary relationships among species. These trees illustrate the evolutionary history of a set of taxa, with branches symbolizing speciation events and nodes indicating common ancestors. While traditional phylogenetic trees are effective in illustrating evolutionary histories for many groups of species, some lineages have more complex evolutionary patterns that simple trees cannot adequately represent (Mallet et al., 2016). In cases where evolutionary histories involve reticulate events such as horizontal gene transfer or hybridization, phylogenetic networks offer a more comprehensive representation. Phylogenetic networks are extensions of phylogenetic trees that can capture conflicts in evolutionary relationships arising due to complex evolutionary processes (Jetten & van Iersel, 2018).

So, the evolutionary history of species can be represented through phylogenetic trees, which depict the relationships and divergence among different taxa. But, how to construct such a phylogenetic tree? Morphological traits have traditionally been used to construct phylogenetic trees, where similarities and differences in physical features are compared to infer evolutionary relationships. This approach assumes that the species gains the trait from its ancestor. On the other hand, we can reconstruct the evolutionary history of more complex patterns, such as the genome of the species. This approach, called molecular evolution, has become increasingly important in phylogenetics due to its ability to provide more precise and detailed information about evolutionary relationships (Shakya et al., 2020).

In the field of phylogenetics, we can distinguish two approaches: the inference of gene trees and the inference of species trees. To infer the evolutionary history of a gene/protein family, we must first find the genes/proteins that belong to the same gene/protein family using homology approaches (**Sections 5A and 5B**). The second step is to build a Multiple Sequence Alignment (MSA), which highlights the conserved positions (bases/amino acids) that came from a common ancestral sequence. From the MSA, different approaches can be used to infer

the phylogenetic tree.

Although the gene trees are of interest on their own, the evolution of the species is important to understand how these genes evolved inside species. To construct a species phylogenetic tree, we can take one universal and unicopy gene (only one copy of the gene in each species genome) and construct its tree or consider multiple genes. Two main approaches are used for multiple genes species tree inference: concatenate genes and supertrees. The concatenation approach, also called Multi-Locus Sequence Analysis (MLSA), involves combining gene sequences from multiple loci into a single supergene alignment, which is then used to infer a species tree (Gadagkar et al., 2005). On the other hand, supertrees are constructed by combining individual gene trees with overlapping taxon sets into a comprehensive phylogenetic tree that includes all taxa from the input trees (Chaudhary et al., 2012). It is possible to construct a species tree using all the genes of the genome, considering the advancements in computational capacities and WGS (Whole Genome Sequencing). However, many studies use only housekeeping genes in the MLSA after proving the capacity of this method to identify and differentiate between closely related bacterial strains (Maiden et al., 1998).

The construction of phylogenetic trees from MSA involves various approaches and tools. The Maximum Parsimony and Maximum Likelihood are the two most used methods in phylogenetic tree construction. Maximum parsimony aims to find the tree that requires the fewest evolutionary changes (number of base substitutions) to explain the observed differences among sequences (Fitch, 1971). On the other hand, maximum likelihood seeks the tree that maximizes the probability of the observed data given an evolutionary model. RAxML and IQ-TREE are the most popular tools for phylogenetic analysis, implementing both maximum parsimony and maximum likelihood algorithms (Nguyen et al., 2015; Stamatakis, 2014). These tools are essential for handling large phylogenomic datasets efficiently, especially for constructing maximum-likelihood phylogenies (Stamatakis, 2014).

6. Objectives of the thesis project

My thesis work was dedicated to studying convergent coevolution through legume-rhizobia symbiosis. In this project, we aimed to answer whether phenotypic convergences arise from similar evolutionary pathways and use identical molecular mechanisms, focusing on nitrogen-fixing symbiosis, an excellent model for studying coevolution and essential ecosystem functions. Benefitting from the genomic resources available in both host legume plants and symbiotic bacteria and generating new datasets, we aimed to answer those questions by combining molecular evolution analysis and functional assays:

- Are similar NCRs also involved in Terminal Bacteroid Differentiation in the under-studied clades?
- Are NCR peptides recruited from the same gene families in IRLC, Dalbergioids, and other clades?
- Have bacterial BacA proteins followed the same evolutionary path for NCR resistance in rhizobia?
- Are plant antimicrobial weapons and bacterial resistance systems coevolving?

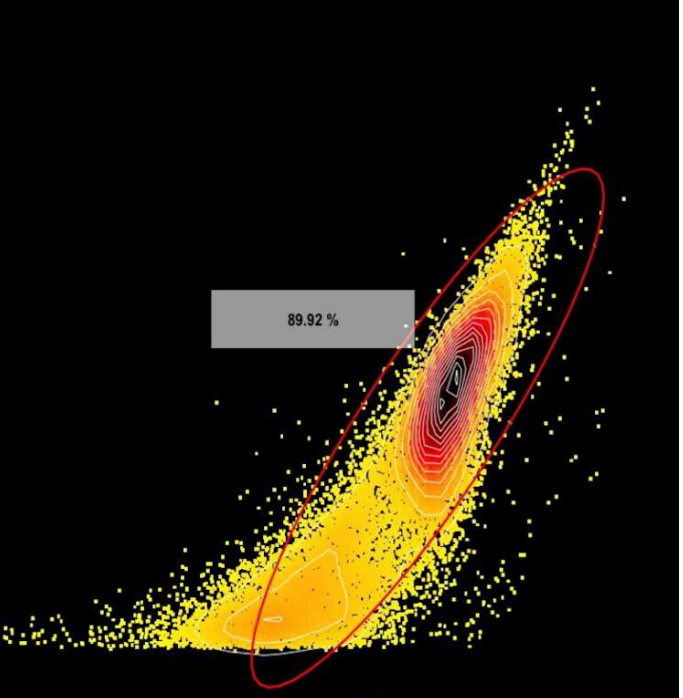
My thesis work was divided into three main objectives to answer those questions. First, we wanted to decipher the evolution of NCR peptides by studying the NCR gene family at different scales. The first scale was the intra-clade comparison of the known NCR peptides within available IRLC legume genomes (*Medicago truncatula, Medicago sativa, Cicer arietinum,* and *Pisum sativum*) and Dalbergioids legume genomes (*Arachis hypogaea* and *Aeschynomene evenia*). The second scale was the inter-clade comparison of NCR peptides, where we searched for new NCRs in all legume species where genomic and nodule transcriptomic data were available, even in legume species that did not undergo the Terminal Bacteroid Differentiation, to see if they are present on the genome or the nodule transcriptome. Once we had all NCR peptides, we did an inter-clade comparison between NCR peptides in different clades, and we tried to elucidate whether convergent evolution occurred and infer the evolutionary history of NCR peptides. In addition to that, to extend our dataset to another unstudied legume clade (Indigofereae) that undergoes Terminal Bacteroid Differentiation, we generated a deep-sequenced nodule and root RNA-seq dataset of *Indigofera argentea* legume to search NCR peptides in this species, and we included them to our inter-clade comparison.

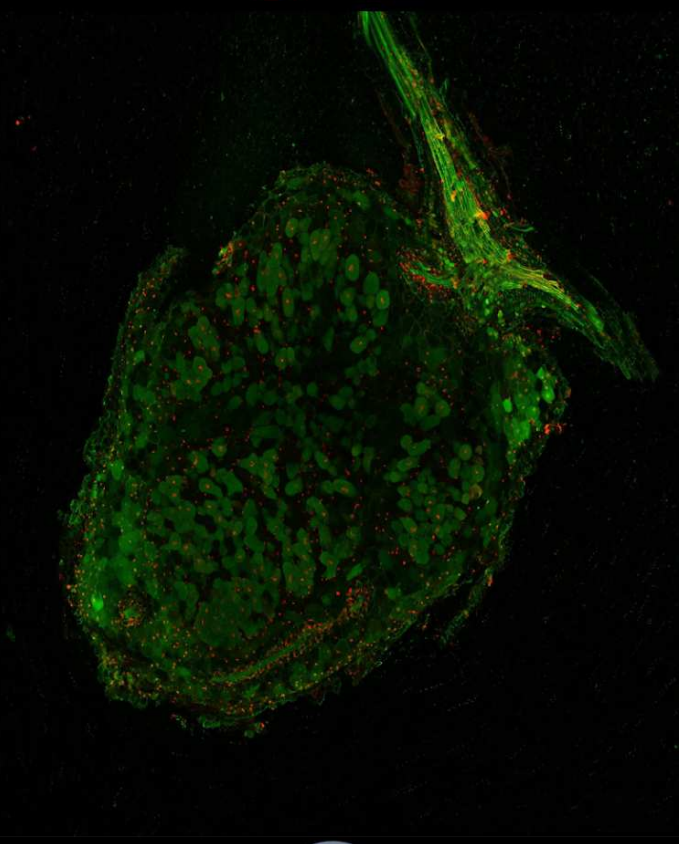
Second, we aimed to resolve the evolutionary history of BacA-like (BacA/BclA) transporters

where BclA transporters in bacterial symbionts of Dalbergioid legumes differ in their sequence from BacA transporters in IRLC symbionts, suggesting that these transporters may also have been repurposed twice independently. However, due to frequent horizontal gene transfers among bacteria, and given that some bacterial genomes encode both BacA and BacA-like transporters, a careful investigation of the evolution of BacA was required. To answer the questions of whether a convergent evolution occurs between BacA and BclA transporters and whether an adaptation takes place in some BacA-like transporters to transport NCR peptides of their corresponding plant hosts, we performed a phylogenetic analysis and network similarity analysis of BacA-like proteins. We also conducted a Multi Locus Sequence Analysis (MLSA) of all bacteria to examine the taxonomic distribution of the BacA-like transporters. Then, to validate that the identified proteins are indeed BacA-like orthologs and to test the specificity of this transporter in symbiosis, we took different BacA-like proteins from the phylogeny and tested *in vitro* and *in planta* if they complement the function of *Sinorhizobium meliloti* BacA which is the transport of NCR peptides.

Finally, we tested functionally the possible coevolution between NCR peptides in legume plants and BacA transporters in rhizobia. Based on the multiscale analysis of NCR peptides, we chose some NCRs to test functionally if NCRs/*S. meliloti* BacA coevolved or not.

During my thesis, I also worked on other projects that are not presented in detail in this manuscript. I collaborated with Jean-Malo Couzigou on the study of the TBD in *Lupinus* species, where I did *in-planta* and flow cytometry experiments to check the endoreduplication of plant cells and bacterial cells in *Lupinus angustifolius* and *Lupinus albus* nodules (Ledermann et al., in preparation). I also worked with Peter Mergaert on the resistance of insect gut symbionts to membrane-damaging antimicrobial peptides called CCRs (Crypt-specific Cysteine-Rich), where I performed *de novo* transcriptome assembly of infected and uninfected *Riptortus pedestris* insect, searched for new CCR peptides using HMM profiles and performed differential expression analysis of all transcripts and CCRs between the two conditions (Lachat et al., 2024). I also collaborated with Besma Bouznif on the identification of peanut symbionts grown on soil samples collected from Tunisia. In this project, I performed the *de novo* assembly and annotation to generate a complete and circularized Oxford Nanopore Technologies (ONT) long read-based genome sequences of five nitrogen-fixing symbionts belonging to the genus *Bradyrhizobium,* collected and characterized by Besma Bouznif (Bouznif et al., 2024). The corresponding articles are in the Annexes section.





II. Results

1. Structural phylogenetics reveals convergent evolution of cysteine-rich peptides in legume-rhizobium symbiosis

A. Foreword

In this project, we studied the evolutionary history of NCR peptides, combining sequence-based, structural phylogenetics, and functional assays. To answer the questions of whether there are other NCR peptides in IRLC and Dalbergioids clades, whether there are NCR peptides in unstudied legume species and clades that undergo TBD, whether there are no NCR peptides in legume species that did not undergo TBD, study their molecular identity and infer their evolutionary history, we performed an inter and intra-comparison of those Nodule-specific Cysteine-Rich antimicrobial peptides.

First, we classified the known NCR peptides from IRLC and Dalbergioids clades using homology, orthology, and clustering approaches. Then, we searched for NCR peptides in all legume species where genome and nodule transcriptome are available. A deep-sequenced nodule and root RNA-seq dataset of the *Indigofera argentea* legume from the *Indigofereae* clade was generated to extend our dataset. Moreover, we generated a high-quality de novo transcriptome assembly of two *Lupinus* species from publicly available raw nodule RNA seq and the generated *I. argentea* data.

Second, because of the divergence of NCR peptide sequences, we predicted the 3D structures of NCR clusters and performed a structural clustering approach to regroup them into superclusters. I also used structural phylogenetics to infer the evolution of NCR peptides. From this analysis, we selected nine NCR peptides from different superclusters, and we tested their function *in vitro*.

This work started at the beginning of my thesis project, where I collected and generated the genomic, transcriptomics, and proteomics data. I carried out all the bioinformatic analysis, the *in-planta* experiments, the flow cytometry, and microscopy experiments. The RNA from *Indigofera argentea* legume was extracted by Benoît Alunni. The functional assays were perfomed by Siva Sankari. The paper is ready to submit and will be submitted soon. The data will be available upon submission.

Structural phylogenetics reveals convergent evolution of cysteine-rich peptides in legume-rhizobium symbiosis

Authors: Amira Boukherissa[1,2], Siva Sankari[3], Tatiana Timchenko[2], Peter Mergaert[2], George C. diCenzo[4], Jacqui A. Shykoff[1,] *, Benoît Alunni[2,5]*, Ricardo C Rodríguez de la Vega[1,]*

Affiliations: [1] Écologie Systématique et Évolution, CNRS, Université Paris-Saclay, AgroParisTech, 91198, Gif-sur-Yvette, France

[2] Institute for Integrative Biology of the Cell, CNRS, CEA, Université Paris-Saclay, 91198, Gif-sur-Yvette, France

[3] Stowers Institute for Medical Research, Kansas City, MO, U.S.A

[4] Department of Biology, Queen's University, Kingston, ON, K7L 3N6, Canada

[5] Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France

* Corresponding authors: Jacqui A. Shykoff (Jacqui.shykoff@universite-paris-saclay.fr), Benoît Alunni (benoit.alunni@inrae.fr), and Ricardo C Rodríguez de la Vega (ricardo.rodriguez-de-la-vega@universite-paris-saclay.fr)

B. Abstract

Legume plants under nitrogen deficiency can perform a symbiotic interaction with atmospheric nitrogen-fixing soil bacteria called rhizobia. In five legume clades, an exploitive strategy has evolved in which rhizobia undergo Terminal Bacteroid Differentiation (TBD), where the bacteria become larger, polyploid, and have a permeabilized membrane. Terminally differentiated bacteria are associated with a higher N_2-fixation and, thus, a higher return on investment to the plant. We know that in several members of the IRLC and the Dalbergioid clades, this differentiation process is triggered by a set of apparently unrelated plant antimicrobial peptides with membrane-damaging activity, known as Nodule-specific Cysteine-Rich (NCR) peptides. However, whether NCR peptides are also implicated in symbiotic TBD in other clades and whether these are evolutionary-related remain unknown. Here, to address the molecular identity of NCR peptides and their evolution in different legume clades, we performed inter and intra-clade comparisons of NCR peptides in four legume clades inducing TBD. First, we collected genomic and proteomic data of species for which NCR peptides are known (1523 NCRs). We then used sequence similarity-based clustering to regroup NCR peptides, resulting in over 400 different NCR clusters, each of which was clade-specific. We obtained Hidden Markov Models for each cluster and used them to predict NCR peptides in 17 legume genomes (6 clades) using a tailored gene prediction pipeline and transcriptome matching. Additionally, we generated deep-sequenced root and nodule RNA-seq data of *Indigofera argentea* (Indigoferoid clade) and reported high-quality transcriptomes of *Lupinus luteus* and *Lupinus mariae-josephae* (Genistoids clade), where NCR peptides were also identified. This resulted in a total of 3710 NCR peptides in species that induce TBD. However, the rapid diversification of NCR peptides that reduce the sequence similarities has highly masked the origin of NCR evolution. We obtained high-confidence structural models for one sequence of each cluster and performed structure-based clustering and phylogenetics, which resulted in 23 superclusters (14 inter-clade and 9 clade-specific) that we represent in a structural distance-based tree. Our study revealed that within each clade, NCR evolution is a mix of divergent and convergent processes. We further chose 9 independently evolved NCRs to test *in vitro* whether they are functional analogs in the context of the symbiosis.

C. Introduction

Legume plants (Fabaceae) have evolved the ability to house symbiotic nitrogen-fixing bacteria in their root nodules. When nitrogen is limited, legume plants can enter a symbiotic interaction with N_2-fixing soil bacteria called rhizobia, an umbrella term including alpha-, and beta-proteobacteria members. During this interaction, the legume plant forms root nodules where rhizobia are housed intracellularly as structures called bacteroids that fix atmospheric nitrogen and transfer ammonia to the plant. In return, legume plants provide these microsymbionts with carbon and other nutrients. This interaction initiates after mutual recognition between the host plant and a compatible bacterial partner involving an exchange of signaling molecules between the two partners (Oldroyd, 2013). First, when grown in nitrogen-deprived soil, the legume plant releases flavonoids to the substrate, thereby attracting rhizobium bacteria and triggering the production of nodulation (Nod) factors by the bacteria. On perceiving these Nod factors, the plant root hairs curl to trap the rhizobia and guide them, via infection threads, toward the incipient nodule. When they reach the cortical cells (Gage, 2004), the rhizobia are released by the infection threads and internalized by the nodule cells. Inside the nodule and, more precisely, inside the subcellular compartment called the symbiosome, rhizobial metabolism is rewired, and they become nitrogen-fixing bacteroids.

Depending on the plant host, bacteroids may remain similar to free-living bacteria, their shape and N2 fixation proficiency being unaltered during the symbiosis (Lamouche, Bonadé-Bottino, et al., 2019; Oono & Denison, 2010). However, in some legume plants like *Medicago truncatula* and relatives belonging to the Inverted Repeat Lacking Clade (IRLC), bacteroids undergo a process called Terminal Bacteroid Differentiation (TBD) (Haag & Mergaert, 2020; Lamouche, Bonadé-Bottino, et al., 2019; Oono & Denison, 2010). Terminally differentiated bacteroids are larger, elongated, or spherical cells that have permeabilized membranes. They undergo endoreduplication and become polyploid while losing their ability to divide but fix N_2 more efficiently, providing a higher return on investment for the plant (Alunni & Gourion, 2016; Mergaert et al., 2006). In Medicago and its relatives, this differentiation process has been shown to be induced by small plant antimicrobial peptides called NCR (Nodule-specific Cysteine-Rich), which are highly expressed in nodules of some legume species that trigger TBD (Pan & Wang, 2017; Van de Velde et al., 2010). These peptides are composed of a signal peptide, which drives their secretion, and a 20 to 50 amino acid-long mature peptide, including 4, 6 (type-1 NCR), or 8 cysteines (type-2 NCR) that form two, three, or four disulfide bridges (Montiel et al., 2017). The mature peptides are highly variable at the sequence level except for

a four or six-cysteine pattern (Mergaert et al., 2003). The number of NCR peptides among legume species varies from 7 (*Glycyrrhiza uralensis*) to 700 (*Medicago truncatula*) (Montiel et al., 2017; Young et al., 2011). According to their isoelectric point, NCR peptides can be classified as cationic, neutral, and anionic. In the few cases studied, cationic NCRs display an antimicrobial activity that permeabilizes bacteroid membranes *in vitro*, while no activity has been shown yet for anionic NCR peptides (Maróti et al., 2011, 2015). Furthermore, NCR peptides have been suggested to provoke a cell cycle switch in symbiosis, where it has been shown that NCR247 can inhibit bacterial cell division by interacting with the FtsZ protein (Farkas et al., 2014). Moreover, it has been shown that NCR247 interacts with other bacterial proteins, such as ribosomal proteins and the chaperonin GroEL (Farkas et al., 2014).

NCR peptides are required for terminal bacteroid differentiation and establishing an effective symbiosis between IRLC legumes and rhizobia (Van de Velde et al., 2010), where they control the bacterial life cycle and other cellular pathways (Roy et al., 2020). These peptides are expressed in waves during different stages of nodule formation and bacteroid differentiation (Guefrachi et al., 2014). Recently, it has been shown that at least four individual NCR peptides are essential for the symbiosis in *Medicago truncatula* (Horváth et al., 2015, 2023; Kim et al., 2015). Furthermore, it has been demonstrated recently that NCR247 is involved in iron homeostasis, where they form complexes with heme moieties, facilitating iron uptake by rhizobia (Sankari et al., 2022).

Rhizobia can tolerate the stress provoked by NCR peptides and prevent membrane damage with the help of specific ABC transporters called BacA or BclA (Glazebrook et al., 1993), which are essential for effective symbiosis involving legumes that trigger TBD (Guefrachi et al., 2015; Haag et al., 2011). BacA is an atypical peptide ABC transporter lacking the ATPase domain. Deletion mutants for the *bacA* gene cannot transport NCR peptides and die in the presence of cationic NCR peptides (Barrière et al., 2017). In addition, *bacA* deletion mutants show multiple sensitivities, including increased resistance to bleomycin and modified membrane composition (Ferguson et al., 2002; Marlow et al., 2009). Furthermore, it has been recently demonstrated that BacA and BclA have been repeatedly co-opted for symbiotic interactions with eukaryotic hosts to cope with antimicrobial peptides and that their functional similarity in symbiosis raised from convergent evolution rather than shared ancestry to cope with NCR peptides (Boukherissa et al., 2024). Though BacA and BclA primarily carry the transport of NCR peptides in rhizobia (alpha and beta proteobacteria), recent results suggest that YejABEF may also contribute to the

import of NCR peptides as *yejAEF* mutants are sensitive to at least one NCR peptide *in vitro* (Nicoud et al., 2021).

Terminal bacteroid differentiation has been observed in five different legume clades (Genistoids, Mirbelioids, IRLC, Indigoferoids, and Dalbergioids) (Oono et al., 2010). The role of plant-secreted NCR peptides in this process is known only in two of them, IRLC and Dalbergioids (Czernic et al., 2015; Montiel et al., 2017). NCR peptides from the IRLC and Dalbergioid clades have different sequences and cysteine motifs, but both induce TBD in the symbiont. Indeed, NCR peptides may have evolved independently in IRLC and Dalbergioid clades, supporting the idea of convergent evolution driving symbiont terminal differentiation (Downie & Kondorosi, 2021). Nevertheless, a recent phylogenetic study between plant defensins and NCR peptides demonstrated that they may share the same origin (Salgado et al., 2022). However, all the studies about NCR peptides were sequence-based and limited to two clades and a small subset of NCR peptides. Therefore, the presence of NCR peptides in other clades, their molecular identity, and their evolution remain unknown. Here, we examine how NCR peptides evolved in legumes and how they are associated with TBD. We combine molecular evolution analysis and functional assays to study the molecular identity and the evolution of NCR peptides at different scales. This study uses a combination of sequence-based, statistical, and structural analyses to analyze the publicly available and newly generated RNA-seq, genomic, and inferred proteomic data of legume species. We report an exhaustive list of NCR peptides of four legume clades that trigger TBD grouped by sequence-based clusters and structure-based superclusters. Although the sequence analysis demonstrates that NCR peptides are clade-specific and highly diverse, the structural analysis grouped NCRs from different clades together and detected a hundred clusters in the same structural supercluster. This suggests that NCR peptides may have evolved from the same family in some clades. Furthermore, the presence of NCR-like genes in legume clades that did not induce TBD suggests that NCR genes evolved from the same gene family but gained the function of inducing TBD independently in some legume clades. This paper highlights the importance of structure-based analysis in the study of the evolution of protein families with highly diverse sequences.

D. Results

1. NCR peptides are clade-specific at the sequence level

To decipher the evolutionary history of NCR peptides, we first studied the known NCR peptides from IRLC and Dalbergioid clades at intra-clade and inter-clade scales. We performed homology, orthology and Markov clustering analysis of all proteins (including NCR peptides) of all legume species where NCR peptides are known and the inferred proteomes are available, i.e. *Medicago truncatula*, *Medicago sativa, Cicer arietinum* and *Pisum sativum,* from IRLC, and *Arachis hypogaea* and *Aeschynomene evenia* from Dalbergioids (see Materials and Methods). This approach led us to the identification of 63,490 orthologous clusters from 483,710 proteins. Among those clusters, 18,856 were inter-clade, and 44,634 were clade-specific (36,792 IRLC clusters and 7,842 Dalbergioid clusters). All NCR-containing orthologous clusters are clade-specific **(Figure S1)**. In the IRLC clade, among the 1523 NCR peptides, 1492 were clustered (assigned to an orthologous group) into 651 clusters. Of them, 203 were clusters containing exclusively NCR peptides (568 NCR peptides), and the remaining 448 clusters (924 NCR peptides) were NCR-mixed orthologous groups with at least one NCR and one no-NCR. Among the 448 NCR-mixed clusters, 238 clusters were NCR-monotypic clusters with only one NCR and at least one other protein. In the Dalbergioid clade, 117 of the 155 NCR peptides were clustered into 20 clusters. Of them, 7 were clusters containing exclusively NCR peptides (40 NCRs), and the remaining 13 clusters (77 NCRs) were NCR-mixed orthologous groups with at least one NCR and one no-NCR, seven of which were NCR-monotypic. One big cluster with 53 sequences represented almost all the *Aeschynomene evenia* NCR peptides, and all the other clusters were small clusters of *Arachis hypogaea* NCRs, which highlights the sequence divergence of NCR peptides even within a clade. Furthermore, all the IRLC NCR clusters contain only NCR peptides with four or six cysteine motifs (hereafter called type-1) in the mature peptides, while 95% of the NCR peptides in the Dalbergioids NCR clusters had a defensin-motif with eight cysteines (type-2). For further analysis, we consider only clusters with at least two NCR peptides with a signal peptide, where we excluded the NCR-monotypic clusters after filtering out the sequences without a signal peptide. Consequently, we end up with 385 NCR IRLC (1191 NCRs) clusters and 11 NCR Dalbergioids clusters (102 NCRs).

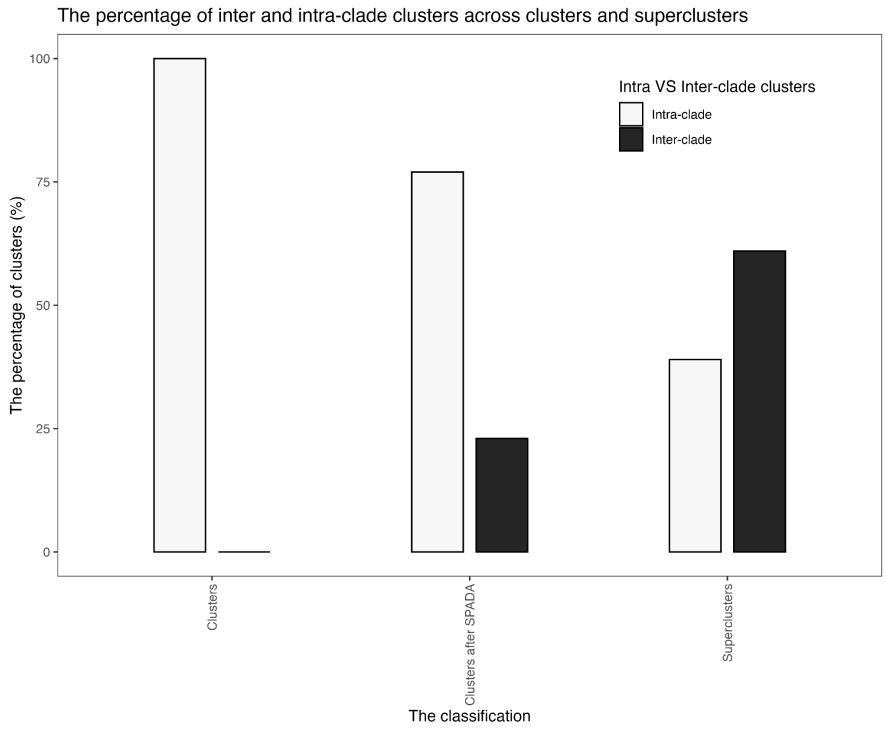


Figure S 1 NCR clusters are clade-specific considering IRLC and Dalbergioids but inter-clade considering four TBD-inducing clades after an exhaustive search with SPADA

NCR peptides have relatively conserved structures despite their sequence divergence, with more than 60% of inter-clade superclusters. The bar plots represent the percentage of clade-specific clusters versus inter-clade clusters in the first computed clusters from IRLC and Dalbergioids (left), in the reconstructed clusters from the recovered NCR peptides with SPADA (middle), and in the superclusters constructed based on 3D structures (right).

2. Newly identified NCR peptides in Indigoferoids and Genistoids clades

In addition to IRLC and Dalbergioids, some Indigoferoids and Genistoids induce TBD (Oono et al., 2010), but their NCR peptides remain undescribed. Therefore, to expand our NCR dataset to other legume clades, we generated deep-sequenced root and nodule RNA-seq data of *Indigofera argentea* legume from the Indigoferoid clade. We obtained a *de novo* transcriptome assembly and annotation of the generated *I. argentea* RNA-seq data and also of two publicly available raw RNA-seq data from nodules of two *Lupinus* species from the Genistoids clade (*L. luteus* and *L. mariae-josephae*) (Keller et al., 2018). In order to confirm that *I. argentea* legume induces TBD, we quantified the DNA content and size of the bacteroids in nodules of *I. argentea* using flow cytometry. This analysis showed an increase in the DNA content with peaks at 3C, albeit barely enlarged bacteroids (**Figure 28a, 28b**). Moreover, the confocal microscopy did not show significant cell enlargement (replicate experiments in progress).

The *de novo* assembly of *Indigofera argentea, Lupinus luteus,* and *Lupinus mariae-josephae* transcriptomes generated 277,022, 156,834, and 152,943 contigs, respectively. Assembled contigs present different splice variants of one gene that we merged into one contig called "supertranscript". The resulting assemblies contain 72,846, 57,642, and 55,700 supertranscripts for *I. argentea, L. luteus,* and *L. mariae-josephae,* respectively. Based on BUSCO genome completeness and the mapping of reads against the de novo assemblies metrics, the assemblies are of high quality, with more than 99% of the cleaned reads mapped to their corresponding contigs, and >85% could be mapped to their corresponding "supertranscript". Moreover, 95% of the Viridiplantae and 86% of the Fabales BUSCO genes were identified as complete and single-copy for the three species (**Figure S2**).

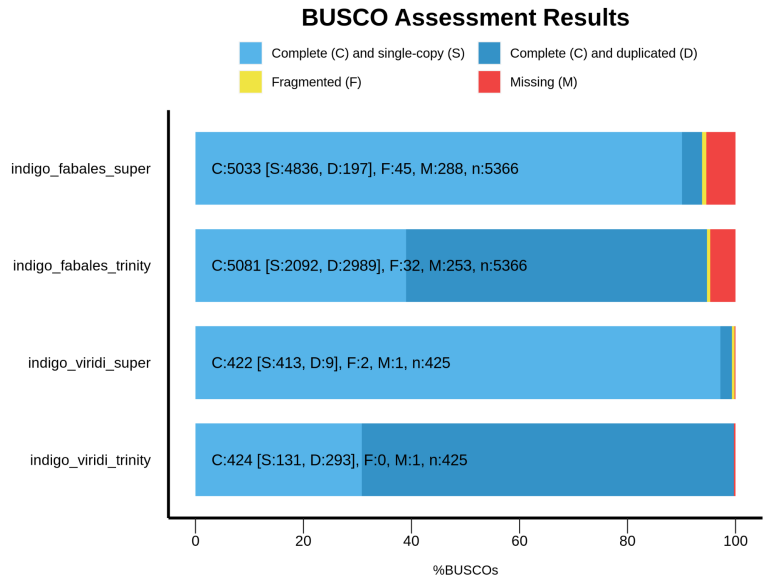


Figure S 2 BUSCO assessment of the completeness of the *de novo* transcriptome and supertranscripts of I. *argentea.*

95% of the Viridiplantae and 86% of the Fabales BUSCO genes were identified as complete and single-copy for the three species.

The search for NCR peptides in these three species was performed using the SPADA pipeline, one run for Genistoids species and two runs for *I. argentea* (see below). A total of 129, 238, and 259 putative NCR peptides were identified in *I. argentea, L. luteus,* and *L. mariae-josephae,* respectively. From these, we annotated 12 (up from 6 recovered in the first SPADA run), 87, and 69, respectively, as "NCR" because they were differentially expressed in nodules, were not longer than 100 amino acids (only for the first SPADA run) and had at least 4 cysteines in the predicted mature peptide. The average and the median lengths of the mature NCR peptides were 38 and 34, respectively, in *L. luteus* and 36 and 33, respectively, in *L. mariae-josephae* (**Figure 29a**). However, in *I. argentea,* the average and the median lengths of the mature NCR peptide

were 81 and 76, respectively, which is higher than expected (**Figure 29a**). Half of the *I. argentea* NCR peptides had an NCR motif with four or six cysteines, and the other half had a defensin motif with eight cysteines. For *L. luteus* and *L. mariae-josephae,* only around 14% of the annotated NCR peptides had a defensin motif.

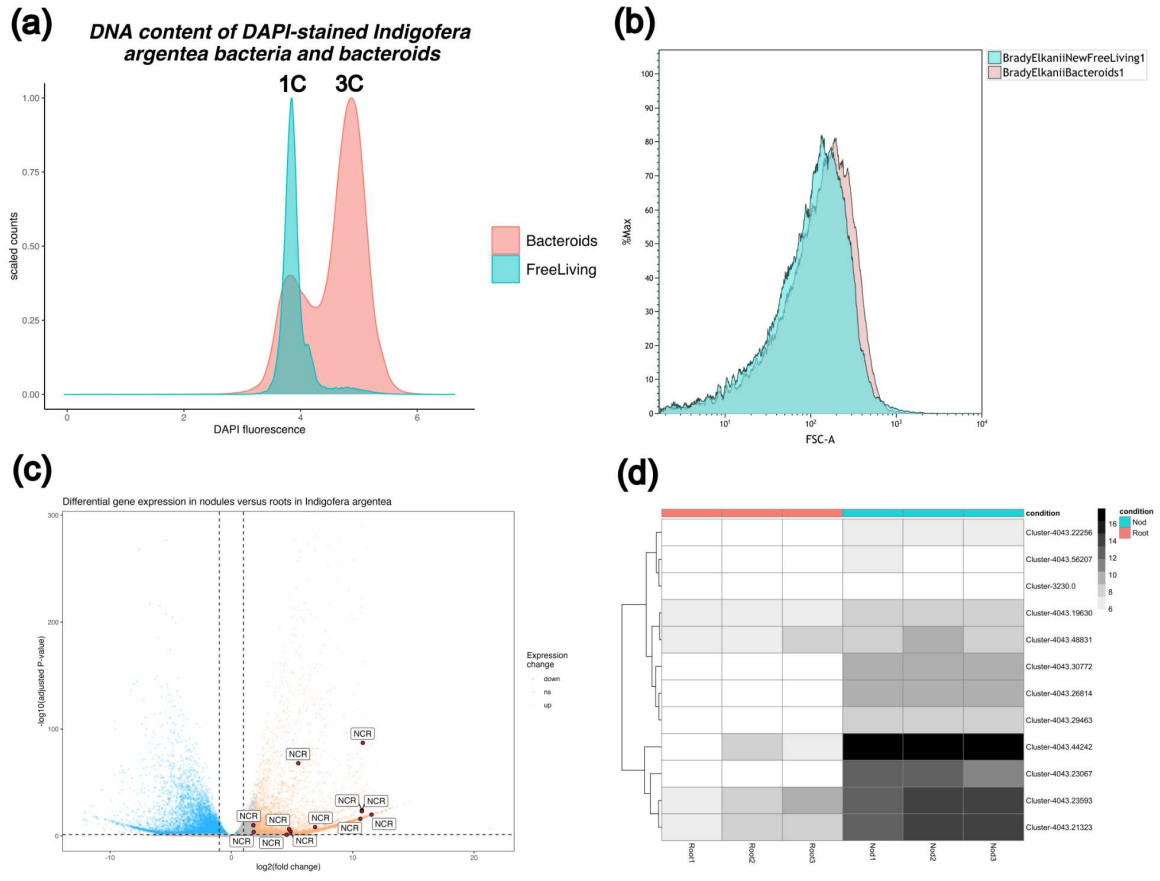


Figure 28 *Indigofera argentea* induces a moderate TBD and expresses NCR peptides.

The few differentially expressed NCR peptides in nodules of *I. argentea* may induce a moderate TBD with only 3C and barely enlarged cells. (a) The DNA content of DAPI-stained *B. elkanii* strain SA281 bacteria and bacteroids isolated from *I. argentea* nodules was measured by flow cytometry, and (b) the size of bacteria and bacteroids. (c) Volcano plot of all *I. argentea* transcripts where the down-regulated transcripts (Log2FoldChange < 0.5) in nodules are in blue (right), the up-regulated transcripts (Log2FoldChange > 2) are in orange (left), the transcripts that had non-significant expression difference between roots and nodules are in grey (middle) and the *NCR* genes among the up-regulated transcripts are annotated and colored in red. (d) Heatmap of the expression of the 12 NCR peptides in the three replicates of nodules (green) and roots (red) of *I. argentea.*

We classified the identified NCRs based on the Hidden Markov Model (HMM) constructed from multiple sequence alignments of 396 sequence-based clusters (385 IRLC and 11 Dalbergioids). We found no matching profile for 11 out of the 12 *I. argentea* NCRs; thus, at the sequence level, *I. argentea* NCRs are also largely clade-specific. In contrast, 75 out of 87

L. luteus and 61 out of 69 *L. mariae-josephae* NCRs matched HMM profiles of IRLC clusters, and one additional *L. luteus* sequence matched the HMM profile of a Dalbergioid NCR cluster. This leaves 11 and 8 of the *L. luteus* and L. *mariae-josephae* NCRs as clade-specific at the sequence level, respectively. According to this sequence-based approach, none of the NCR peptides previously identified as important in *Medicago truncatula* - *Sinorhizobium meliloti* symbiosis (NCR247, NCR211, NCR169, NCR-new35, and NCR343) (Horváth et al., 2015, 2023) would have homologs in either *I. argentea, L. luteus* or *L. mariae-josephae.*

The differential expression analysis showed 13540 up-regulated and 12356 down-regulated genes in the nodules compared to the roots. As expected, the 12 genes encoding NCR peptides in *I. argentea* are highly and differentially expressed on nodules (**Figure 28c, 28d**), and one of them belongs to the 10 most abundant transcripts. Additionally, three of the four NCR peptides most differentially expressed on nodules (clustering together in the heatmap) (**Figure 28d**) are cationic. Furthermore, the transcript abundances calculated using TPM (Transcripts Per Million) between the three *Lupinus* species show a significantly lower expression of NCR peptides in *Lupinus albus* than the two other *Lupinus* species. Indeed, L. *albus* has fewer NCR peptides than the two others.

3. Distribution of NCR peptides across legume species

Once our dataset was expanded to other clades, we also expanded our NCR dataset inside each clade and each species, where we searched for new NCR peptides in all legume species where genomic and nodule transcriptomic data were available (14 in total) using SPADA that took as input the HMM profiles built from our IRLC and Dalbergioids NCR clusters and the CRP (Cysteine-Rich Peptides) clusters (see Materials and Methods). This analysis allowed us to recover NCR peptides in 14 legume species, including 4 species for which NCR repertoires have never been described and novel NCRs in well-studied clades (e.g. 13% to 70% of NCRs in six IRLC species were newly identified here) (**Figure 29a**). While almost all newly identified NCR peptides in IRLC and Genistoids were classified with our known NCR clusters, in the Dalbergioids, only 24 to 32% of NCRs were classified, and only one of the recovered Indigoferoid NCRs (by the second SPADA run) was classified with one Dalbergioid cluster. Additionally, the six *I. argentea* NCRs recovered in the first SPADA run match a profile constructed with CRP. A second round of SPADA, using HMM profiles built from our expanded NCR clusters (including the newly identified NCRs from IRLC, Dalbergioids, and Genistoids) and supplemented with one *I. argentea* NCRs HMM from (Ren, 2018), recovered

six new *I. argentea* NCRs, five matching the *I. argentea* profile and one matching a profile from a Dalbergioid cluster.

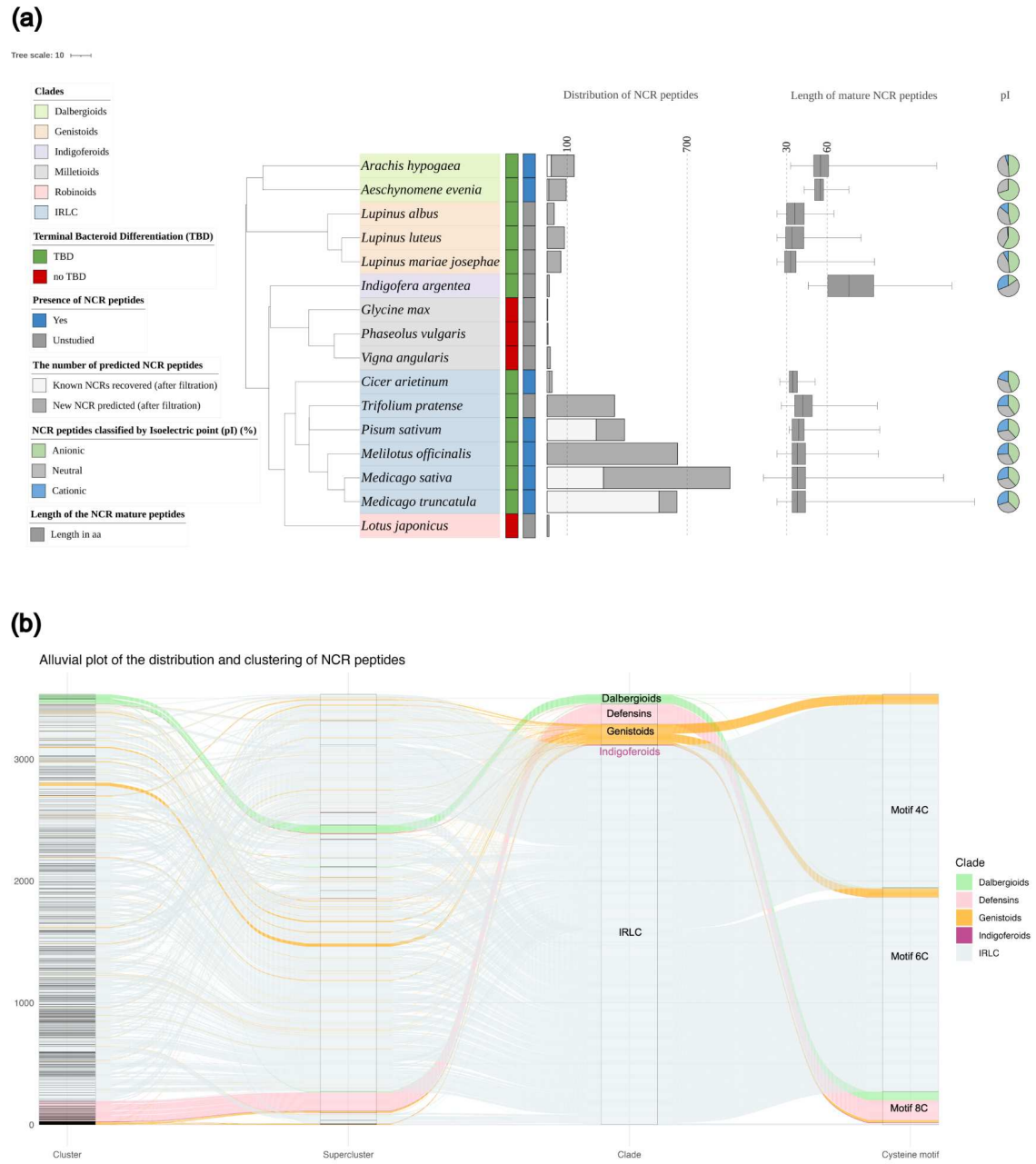


Figure 29 Distribution and characteristics of NCR peptides across the legume phylogeny and across NCR clusters and superclusters.

(a) Legume species tree generated by TimeTree (ref) where we annotated the number of the recovered known NCR peptides and the new NCRs from the predicted NCR peptides after filtration, the length of their mature peptides and the percentage of anionic (pI <= 6), cationic (pI >= 8) and neutral (6 < pI < 8), where NCR peptides are present in all legume clades that induce TBD. (b) Alluvial (RiverPlot) of all NCR peptides regrouped by sequence-based clusters, structure-based superclusters, clades, and cysteine motifs where we can trace the fate of each NCR or cluster. For instance, IRLC NCRs regroup with Genistoids in the same clusters and superclusters, while Dalbergioids and Indigoferoids regroup with defensins in the same superclusters.

The average length of the mature NCR peptides was 34 for IRLC and 40 for Genistoids. The mean lengths of the Dalbergioids and Indigoferoids mature NCRs were 55 and 81 aa, respectively (**Figure 29a**). As expected, the anionic NCR peptides are the most abundant in all the studied legume species, except for *I. argentea*, where the neutral NCRs are the most abundant (**Figure 29a**). *Astragalus sinicus* from the IRLC clade had the highest percentage of cationic NCR peptides at 35%, while *Aeschynomene evenia* had the lowest rate with no cationic NCR peptides (**Figure 29a**). Genistoids and the Dalbergioids clades had few to no cationic NCR peptides.

Furthermore, to try to reveal the evolutionary connections between the NCR clusters, we constructed a sequence similarity network analysis using CLANS (CLuster ANalysis of Sequences) of one sequence per cluster (**Figure 30a**). This analysis allows us to separate between defensins, Dalbergioids, and IRLC-Genistoids, which seem to be highly related (**Figure 30a**). However, the presence of one Dalbergioid cluster in the defensins shared cluster and the divergence of amino acid sequences of NCR peptides suggest that some evolutionary connections may be masked by the sequence dissimilarities between NCR peptides from different species and clades.

4. Structural phylogenetics analysis reveals the evolution of NCR peptides

NCR peptides are small peptides (30-50 aa in the mature peptide, with few exceptions) and are divergent at the sequence level. This rapid diversification of NCR peptides that reduce the sequence similarities has likely hidden the evolutionary origin of these peptides and made the inference of their evolutionary history using traditional phylogenetic analysis very difficult. Therefore, in order to gain more insights into the evolution of NCR peptides, we used structural clustering and phylogenetics to study the NCR peptides at the 3D structure level, to see whether they are also divergent at the structure level and to reduce the number of clusters, regrouping them into structural superclusters.

To perform structural analysis, we predicted the 3D structures of 396 NCR peptides (see Material and Methods) and 48 defensins (outgroup) from four legume clades that induce TBD using Alphafold2 (Jumper et al., 2021). After filtering out the structures that had a pLDDT score < 70, a total of 390 NCRs and 48 defensins were kept. The 3D structures of the unclassified clade-specific NCR peptides from *I. argentea* and the *Lupinus* species (Genistoids) were also predicted using AlphaFold2. We excluded five *I. argentea* NCRs as we could not predict their 3D structure with a confident pLDDT score (pLDDT<55). We used Foldseek

(Barrio-Hernandez et al., 2023) to regroup those structures into 23 superclusters, nine of which were small clade-specific clusters, and the remaining 12 were inter-clade. For instance, this analysis allowed us to regroup the cationic clusters of IRLC and Genistoids in the same supercluster (SC156), including the NCR343 that was identified as essential for an effective symbiosis in *M. truncatula.* Defensins regroup together in one supercluster with a few Genistoids, Dalbergioids, and one of the most abundant Indigoiferoids NCRs (**Figure 31, 29b**). The well-studied peptide NCR247 from *M. truncatula* belongs to a depauperate supercluster composed of just 5 sequences, exclusively found in *Medicago* species (**Figure 30a, S3**).

The comparison between the TM scores and the sequence identities inside the biggest supercluster (**Figure 30b**) showed that the 3D structures of NCR peptides are relatively conserved despite the divergence of their amino acid sequences, which justified why the evolutionary connections were hidden using sequence-based approaches. Moreover, the comparison between the TM scores inside one supercluster and between one supercluster and all the others showed high differences between TM scores inside the supercluster of Dalbergioids compared with all others and inside the supercluster of Defensins compared with all others (**Figure 30c**), while the TM scores inside one IRLC-Genistoids supercluster and outside it showed bi-modal distribution for the outside distribution, one remote peak for TM scores with Dalbergioids-defensins superclusters and one peak close to overlapping the inside distribution that represents the TM scores with other IRLC-Genistoids superclusers (**Figure 30c**). This analysis highlights the two different evolutionary trajectories of defensins-Indigoferoids-Dalbergioids and IRLC-Genistoids NCR peptides.

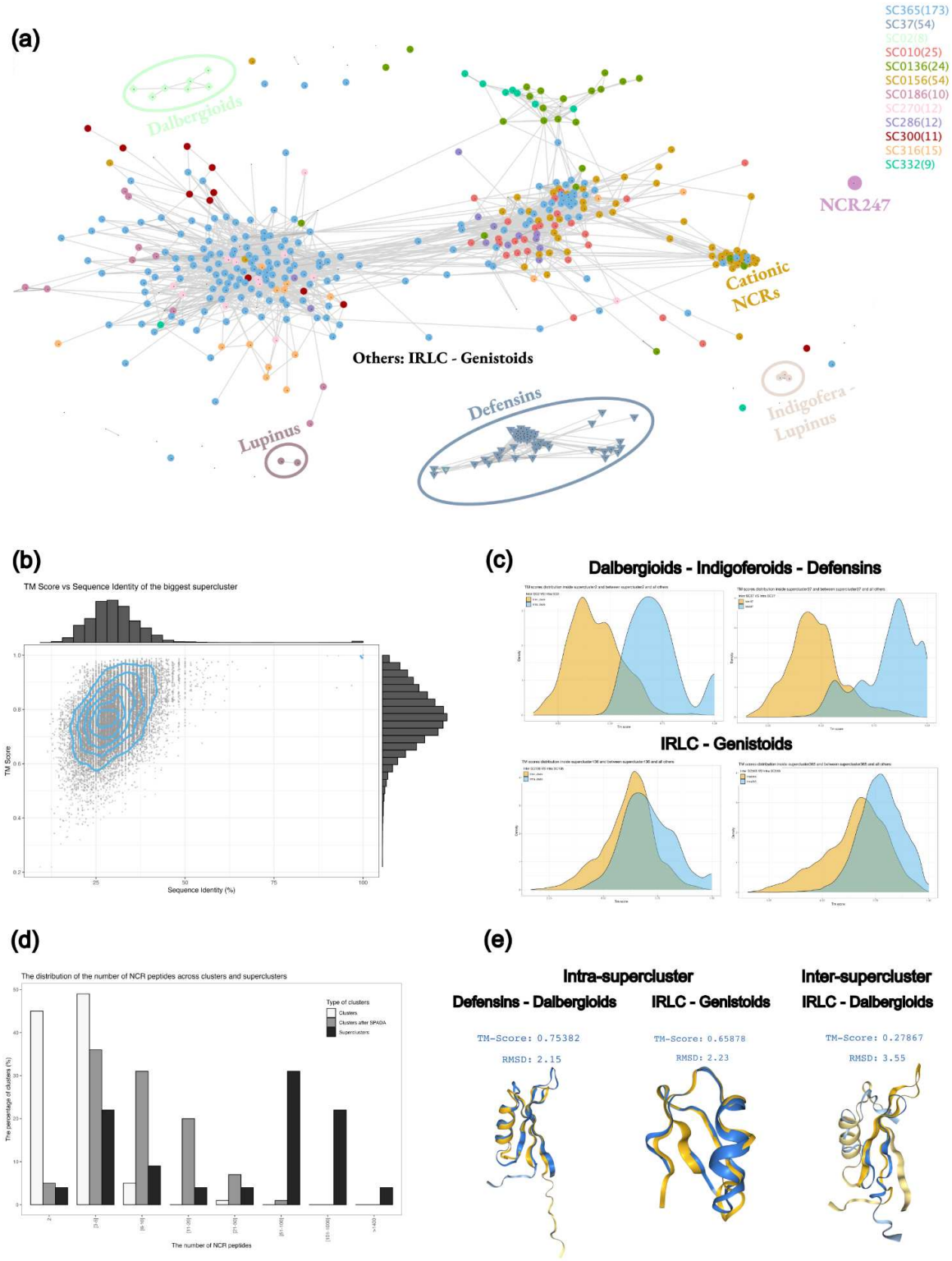


Figure 30 Structural conservation of NCR peptides displaying a high level of sequence divergence.

(a) CLANS (Frickey & Lupas, 2004) sequence similarity network of the same sequences used to perform structural analysis (one per cluster + monotypic). The colors are the same used to represent the superclusters in the structure-based phylogeny below, where we see that only IRLC and Genistoids had related sequences. (b) Dot plots and

histograms that represent the sequence identity between sequences and the TM scores between structures of NCR peptides of the biggest supercluster where we demonstrated that inside one supercluster, the sequences are divergent (sequence identity around 27%). However, their structures are conserved (TM score around 0.75). (c) Comparison between TM scores inside one supercluster (blue) and between one supercluster and all the others (yellow) where two different profiles are observed. The first is when we compare Dalbergioids (or defensins) with all other superclusters (IRLC-Genistoids), where the TM scores, in this case, are weak (density yellow in up plots). The second is when we compare IRLC-Genistoids superclusters with all others (other IRLC-Genistoids and only 2 superclusters of defensins-Dalbergioids), the TM scores in this case are high (almost overlapping the TM scores inside the supercluster) indicating the conservation and homology of IRLC and Genistoids NCRs even in other superclusters. (d) Bar plots of the percentage of NCR clustering according to the number of NCR peptides in the cluster where small clusters are abundant when we consider only IRLC and Dalbergioids NCRs (clade-specific), moderate when we consider four legume clades and more NCR peptides and big superclusters when we classify NCRs by their structures. (e) Structural alignments of intra-supercluster structures with high TM scores versus inter-superclusters with low TM scores.

In order to decipher the evolution of NCR peptides and defensins and have a better overview of our superclusters, we used Foldtree (Moi et al., 2023). This structural phylogenetic approach uses structural distances to build a tree (see Material and Methods). This analysis allowed us to build structure-based trees of all our structures and for each supercluster (**Figure 31**). These structural-based trees were supported by structural alignments of each supercluster (see Material and Methods). Consistent with the presence of few Dalbergioids and Indigoferoids in the defensins supercluster (**Figure 29b**), the structural phylogenetic tree regrouped together the superclusters of defensins and Dalbergioids, separately from the other IRLC-Genistoids superclusters (**Figure 31**). Additionally, the structural alignment of the defensins supercluster is similar to the Dalbergioids one (**Figure 30e, 31, S3)**, both different from the structural alignments of the IRLC-Genistoids superclusters (**Figure 30e, 31, S3)**. Albeit not fully resolved, two stories emerged from these results, revealing the hidden evolutionary history of NCRs. On one hand, NCR peptides from Dalbergioids and Indigoferoids evolved from defensins. On the other hand, taking into consideration that IRLC and Genistoids are relatively distant clades, the NCR peptides in these two clades are recruited independently by convergent evolution and were then expanded and rapidly diversified in the IRLC clade. Indeed, the NCR peptides in these two clades regroup in the same clusters and superclusters with some species-specific NCR peptides.

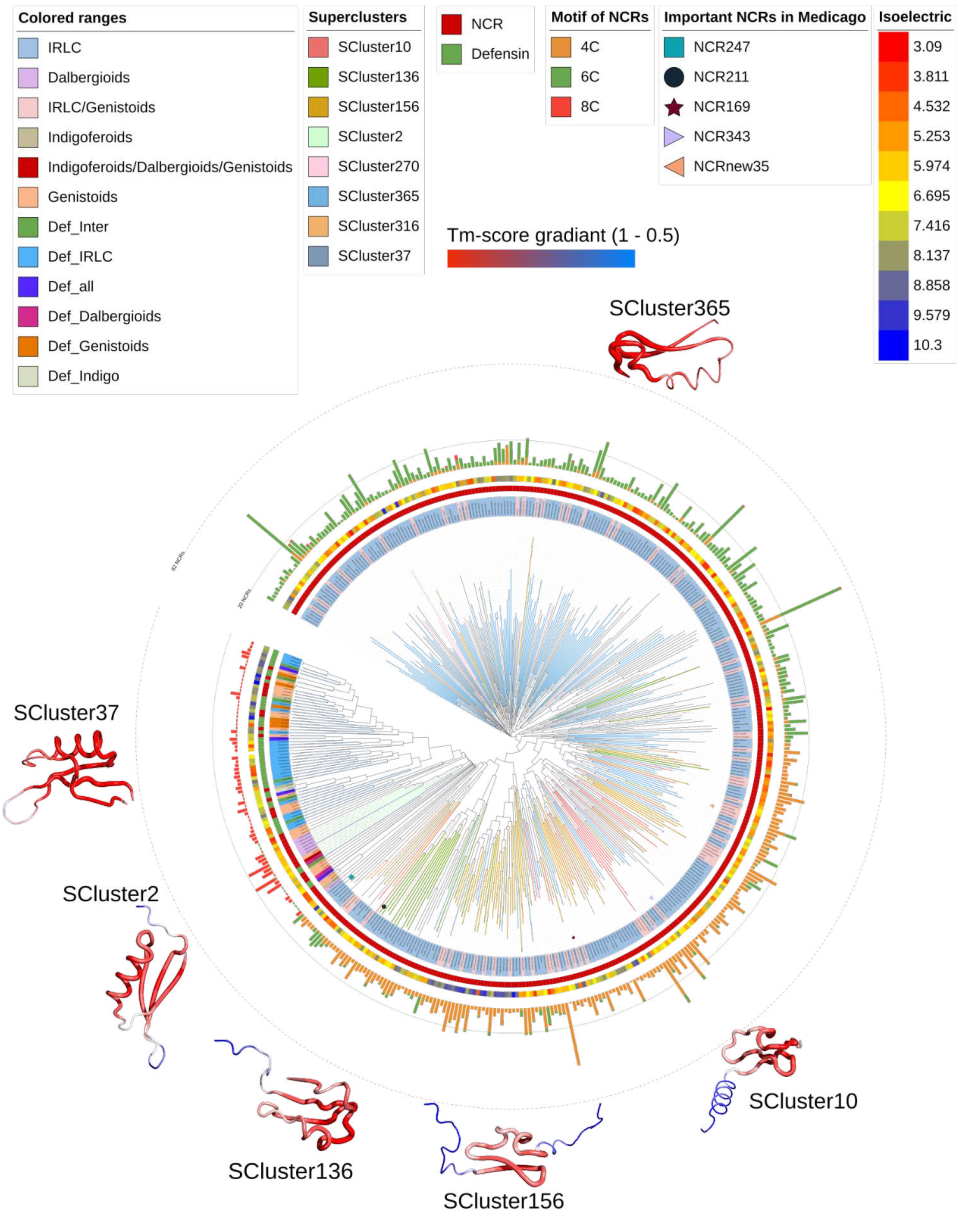


Figure 31 Structural phylogenetic analysis of NCR peptides across legumes

Phylogeny of 444 3D structures of NCR peptides (one per cluster) and defensins produced by Foldtree (Moi et al., 2023), where the branches are colored according to the superclusters defined with Foldseek (ref). Thus, the labels represent the NCR clusters. The labels are colored according to the legume clades. In the first strip, the red represents NCR peptides, and the green defensins. The second strip represents the isoelectric point gradient calculated from the mean isoelectric points of all NCR peptides of each cluster. The multibar plots represent the size of clusters (number of NCR peptides), where the red represents the number of NCR peptides with 8C motif, green with 6C motif, and orange NCRs with 4C motif. The structural alignments represented by a sausage representation are the alignments of all the structures of the supercluster with a color gradient from red (TM score 1) to blue (TM score 0.5). The identified NCR peptides important for an effective symbiosis in *M. truncatula* are annotated in the branches with different motifs.

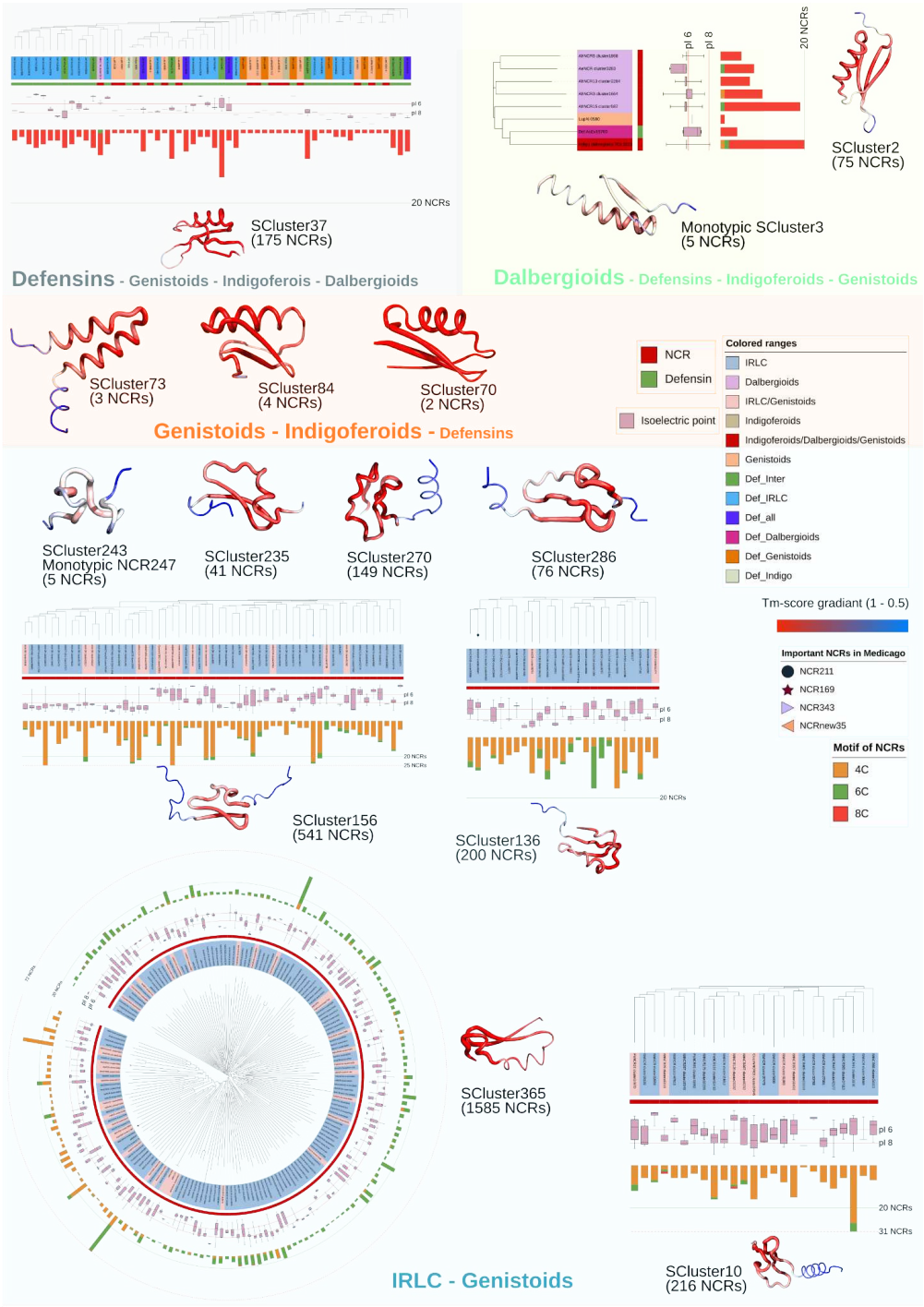


Figure S 3 Structural phylogenetic analysis of each structural supercluster across legumes.

Structure-based Foldtree phylogenetic trees of the most representative superclusters and structural alignments of these superclusters and other small superclusters. In each tree, the labels are colored according to the clade. The box plots represent the isoelectric point distributions. The multibar plots represent the size of clusters (number of NCR peptides), where the red represents the number of NCR peptides with 8C motif, green with 6C motif, and orange NCRs with 4C motif. The structural alignments represented by a sausage representation are the alignments of all the structures of the supercluster with a color gradient from red (TM score 1) to blue (TM score 0.5). The identified NCR peptides important for an effective symbiosis in *M. truncatula* (if present in presented superclusters) are annotated in the branches with different motifs.

5. Function of newly-found NCR peptides *in vitro*

In order to validate the approach used to predict the NCR peptides, test their functions, and, most importantly, to validate that the NCRs that regroup with defensins are truly NCRs, we synthesized nine evolutionary distant NCR peptides from different clades and different superclusters (**Table S1**). We selected two cationic *C. arietinum* (IRLC) NCRs from the biggest NCR cluster belonging to the biggest supercluster, one predicted and one previously known. One highly expressed neutral *L. luteus* (Genistoid) NCR from the second-biggest cluster was chosen. We picked two cationic NCRs from the third biggest cluster, one from *Astragalus sinicus* and one from *T. pratense* (IRLC). Moreover, we have chosen three NCR peptides from the cationic supercluster: one *M. sativa,* one *M. truncatula,* and one *T. pratense* (IRLC). Furthermore, we selected one highly expressed and cationic *I. argentea* (Indigoferoid) NCR peptide belonging to the defensins supercluster. The mass of these peptides ranged from 3.23 to 5.75 kDa (**Table S1**).

Peptide #	Peptide Name	Clade	Supercluster	Isoelectric point	Molecular Weight (kDa)	Ext. Coefficient (e/1000)
1	AsNCR100	IRLC	356	9.30	4.72	1.49
2	Incrp0000	Indigoferoids	37	8.4	5.75	-
4	LuCluster1026	Genistoids	356	6.9	3.23	5.50
6	MsCluster27085	IRLC	156	9.68	3.90	-
7	MtCluster30202	IRLC	156	9.66	4.15	8.48
8	TpCluster1457	Genistoids	356	10.31	4.22	-
9	TpCRP1190	IRLC	156	10.27	4.30	1.49
10	Cacluster155	IRLC	365	9.94	3.74	2.98
11	CaNCR63	IRLC	365	10.04	4.13	2.98
Control	NCR247	IRLC	Monotypic	9.5	3.00	2.98

Table S 1 NCR peptides selected for *in vitro* functional assays.

The table indicates for each selected NCR its clade, supercluster, isoelectric point, molecular weight, and its extinction coefficient.

It has been reported recently that the peptide NCR247 can bind to and sequester heme moieties, facilitating the import of iron by the rhizobial symbiont (ref). To check if other NCR peptides could also bind to the heme, we measured the UV-Vis absorption spectrum of 1:1 heme-bound NCR peptides. Interestingly, among the nine tested NCRs, only one *M. truncatula* bound heme with an absorption maximum at 420 nm (**Figure 32a**). This is different from the *M. truncatula*

peptide NCR247, which bound haem with absorption maxima at 360 and 450 nm. However, the *M. sativa* NCR from the same supercluster did not bind to haem (**Figure 32a**). Moreover, the *A. sinicus* NCR peptide from another supercluster bound heme also at 420 nm but to a lesser extent (**Figure 32a**). NCR peptides are known to act as antimicrobials against a variety of bacteria. We checked the anti-bacterial activity against *S. meliloti* and an unrelated gamma-proteobacterium *E. coli*. As expected, the neutral NCR peptide from *L. luteus* did not exhibit antimicrobial activity against *S. meliloti*, while 7 of the 8 cationic NCR peptides inhibited *S. meliloti* growth. These seven NCR peptides are toxic to *E. coli* (**Figure 32b, d**) but to various extents. We then tested if these peptides could induce ploidy since it is a characteristic feature of differentiated bacteroids. Ploidy is measured by directly measuring the DNA content of synchronized, peptide-treated cells through flow cytometry. At lower peptide concentrations, polyploidy is induced by 7 of the 8 cationic peptides. The neutral *L. luteus* NCR peptide did not induce ploidy abnormalities (**Figure 32c**), yet it is differentially and highly expressed in nodules; we suggest that this NCR peptide is essential in the presence of other NCR peptides. The cationic *T. pratense* NCR peptide displayed less change in the DNA content of *S. meliloti* than the other peptides (less than 10% of polyploid cells) (**Figure 32c**). Notably, the *I. argentea* NCR peptide belonging to the defensin supercluster with a defensin motif (eight cysteines) induced genomic DNA amplification (**Figure 32c**), another feature of NCR peptides after being highly and differentially expressed in nodules. This result validates our approach to searching and classifying NCR peptides and supports the suggestion of the evolution of Indigoferoids NCR peptides from defensins.

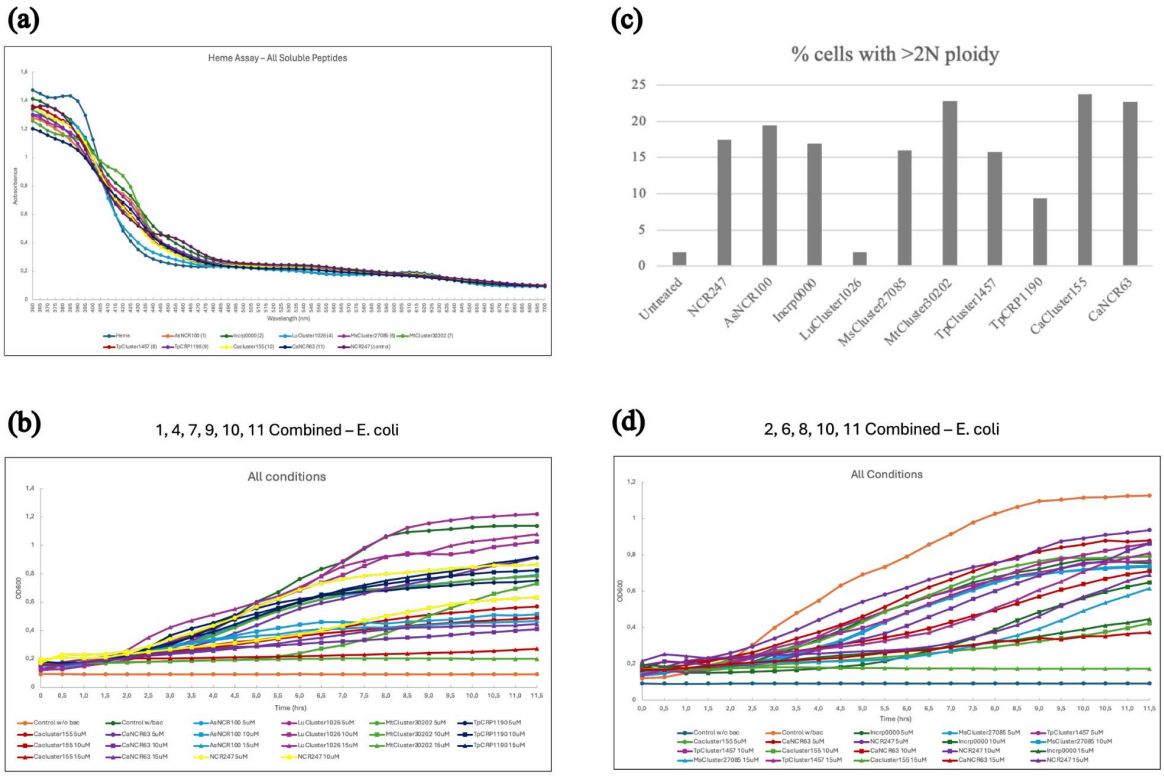


Figure 32 Seven of the nine selected NCR peptides induced TBD features on free-living bacteria *in vitro*.

(a) The absorptions of the heme-bound NCR according to the wavelengths (nm) of each NCR peptide. NCR247 and heme were used as controls. (b) The percentage of cells of *S. meliloti* bacteria treated with one NCR peptide that showed ploidy (more than 2C of DNA) of each NCR peptide. Untreated *S. meliloti* bacteria were used as control. (b) and (d) The toxicity of NCR peptides to *E. coli* measured by the growth (OD_{600nm}) of NCR-treated *E. coli* bacteria according to the time. Untreated wild-type *E. coli* and NCR247-treated *E. coli* were used as controls.

E. Discussion

With the development of high-throughput sequencing technologies, the increasing availability of legume genomes and transcriptomes, and the significant progress in structural bioinformatics, we gained new insights into the evolution of NCR peptides in legume species. The use of NCR peptides to induce TBD has been identified only in two of the five clades that induce TBD (IRLC and Dalbergioids), with two different hypotheses of evolution: independent evolution (Czernic et al., 2015) and evolution from defensins (Salgado et al., 2022). Nevertheless, our data and analysis allowed us to identify NCR peptides in two other clades that induce TBD and to decipher the evolutionary history of NCR peptides.

Primarily, our sequence-based study revealed clade-specific and species-specific NCR peptides, highlighting the divergence of amino acid sequences of known NCR peptides even inside the clade and in the same species. These results suggest an independent evolutionary

origin of IRLC and Dalbergioid NCRs (Czernic et al., 2015) and a common origin in the same clade with some species-specific NCR peptides (Montiel et al., 2017). For example, the *Aeschynomene evenia* and *Cicer arietinum* NCRs regroup separately in their clusters, respectively. Moreover, it was previously shown (Raul et al., 2021) that *Arachis hypogaea* has only type-2 (defensin-like) NCR peptides with eight cysteines. However, here, we found 115 new *A. hypogaea* NCR peptides, 25 of which are type-1 NCRs (i.e. with four or six cysteines). Type-2 NCR peptides were more highly expressed in the nodules than the type-1 NCRs. However, studying the sequence variation of NCR peptides was not sufficient for deciphering their evolutionary history because they are highly diverse short proteins. Indeed, different amino acid sequences do not mean that the proteins are different and evolved from different ancestors. Moreover, only two clades among five that undergo TBD were analyzed before.

The orthology and clustering analysis of IRLC and Dalbergioids proteins demonstrated that NCR peptides are clade-specific at the sequence level. However, the analysis of NCR peptides in a broader spectrum, expanding our dataset to four legume clades that undergo TBD and 3710 NCRs, allowed us to identify inter-clade NCR peptides, where NCR peptides from the Genistoids clade regroup with IRLC NCR peptides in the same clusters.

Moreover, our 396 IRLC and Dalbergioid NCR clusters do not contain all the known NCR peptides because we only took the NCRs that belong to an NCR ortholog cluster with at least 2 NCR peptides. One NCR peptide is included in our clusters if it has at least one NCR ortholog (at least two NCRs per cluster). Therefore, in addition to the identification of inter-clade NCR clusters (**Figure S1**) and novel NCR peptides in all studied legume species, profile-based gene calling with SPADA allowed us to capture the NCRs with no orthologs and the monotypic NCR peptides and add them to our clusters. Moreover, a total of 145 *Astragalus sinicus* NCR peptides from (Wei et al., 2022) were kept after the filtration steps, from which 128 were classified and added to our IRLC NCR clusters. With this analysis, we expanded our NCR clusters from two clades to four clades and from 1293 to 3710 NCR peptides.

Interestingly, the widely used NCR in *Medicago truncatula* NCR247 has homologs only in *Medicago sativa*. Recently, a study that identified the NCR peptides from the transcriptome of *Medicago sativa* and *Melilotus officinalis* (Huang et al., 2022) using SPADA reported that no homolog of the NCR peptides known to be essential for symbiosis in *Medicago truncatula* (ie. NCR211 and NCR169) were found in *Medicago sativa* and *Melilotus officinalis*. However, with our analysis, among the newly identified 633 *Medicago sativa* NCR peptides, we found the

homologs of NCR211 and NCR169. Moreover, we also recovered in *Medicago sativa* and *Melilotus officinalis* homologs of the recently identified essential NCR peptides in *Medicago truncatula* (Horváth et al., 2023) (NCR343 and NCR-new35). However, our sequence-based approach identified no homologs of these five essential NCRs in *Medicago truncatula* in any of the other clades.

The differential expression analysis between the three *Lupinus* species validates that TBD correlates positively with the number and expression of NCR peptides, previously suggested in IRLC species (Montiel et al., 2017). On one extreme, *Medicago truncatula* has more than 700 NCR peptides and induces larger alterations to their symbiont, while in the other extreme, *I. argentea*, with only 12 NCRs, induced limited TBD.

The Foldseek superclusters and the Foldtree structure-based phylogeny of NCR peptides elucidated important evolutionary features of the diversification of NCR peptides that could not be resolved with sequence-based approaches. Despite the genomic variation in NCR peptides and defensins, their 3D structures are relatively conserved.

The heme-binding assay results, where only one NCR peptide from *Medicago truncatula* binds well, still in different wavelengths, confirm the rapid divergence of NCR peptides inside the species (Nallu et al., 2014) and suggest a functional divergence in addition to the sequence divergence. Moreover, the two NCRs from *C. arietinum* selected from the biggest cluster, one previously known, and one predicted with our analysis, displayed similar functional features (no heme-binding, same toxicity to *E. coli*), which underscores the efficiency of our approach to predict and classify NCR peptides.

Together, these results suggested that sequence-unrelated structurally similar NCR peptides from relatively distant legume species (Dalbergioids-Indigoferoids) have evolved repeatedly from defensins. On the other hand, the sequence and structure-related similar NCR peptides from relatively distant IRLC and Genistoids species have evolved repeatedly by convergent evolution and then expanded probably by local duplication followed by rapid diversification where they lose their sequence identity yielding to species-specific NCR peptides.

Finally, through this study, we demonstrated how this method could reveal the evolution of NCR peptides hidden by their sequence divergence. Still, the ancestry gene family of NCR peptides from IRLC-Genistoids remains unknown and needs more investigation.

F. Material and methods

1. Bacterial strains, nodulation assays, and analysis

Sinorhizobium meliloti 1021 and *Bradyrhizobium elkanii* SA281 strains were grown at 28°C in YEB (0.5% beef extract, 0.1% yeast extract, 0.5% peptone, 0.5% sucrose, 0.04% $MgSO_4$ $7H_2O$, pH 7.5) or LBMC (LB medium supplemented with 2.5 mM $CaCl_2$ and 2.5 mM $MgSO_4$ and YM (Vincent, 1970) medium, respectively, in the presence of streptomycin (Sm; 500 µg/mL).

Seeds of *Indigofera argentea,* from an accession originally collected in 2010 in the Jizan desert in Saudi Arabia, which is part of the Nagoya protocol since 2020, were provided by Ton Bisseling. Seeds were treated with 96% sulfuric acid for seven minutes before being rinsed six times with double-distilled water. The seeds were then surface sterilized with 4% commercial bleach for 10 minutes and rinsed seven times before being soaked in sterile double-distilled water for three hours at room temperature in the dark. The sterilized seeds were plated on water agar in 9 cm plates and incubated at 4°C for 12 hours in the dark and at 28°C for 24 hours in the dark. After that, the seeds were exposed to light for 4-5 days, and then the germinated seeds were planted in perlite/sand (2:1 vol/vol) humidified with nutrient water in 1.5L pots in the greenhouse (28°C, 16 hours of light and 8 hours of dark, humidity 60%). The seedlings were grown in the greenhouse for 3-4 days without watering and then inoculated with 20mL per pot of *B. elkanii* SA281 at OD_{600nm} of 0.05 and grown for another 2-3 days without watering. The plants were watered every three days, alternating tap water and a commercial N-free fertilizer (Plant Prod solution [N-P-K, 0-15-40; Fertile] at 1 g per liter) (Kazmierczak et al., 2017). Root nodules from inoculated plants and roots from uninoculated plants were collected 8 weeks post-inoculation, immediately frozen with liquid N_2, and stored at -80°C until use. We kept some fresh nodules (non-frozen) for confocal microscopy and flow cytometry experiments.

2. RNA extraction and sequencing

The total RNA from three biological replicates of frozen root and nodule tissue of *Indigofera argentea* was extracted using MasterPure Complete DNA and RNA purification kit following the manufacturer's protocol. We used a turbo DNA-free kit from Ambion, treating 1µg of RNA per reaction to degrade any contaminating DNA. The concentrations of the purified RNA samples were measured using the RNA method using the DeNovix Spectrophotometer DS-11. RNA integrity was assessed with gel electrophoresis.

Library preparation and Illumina sequencing were performed at "Plateforme de Séquençage Haut Débit I2BC" (Gif-sur-Yvette, France). Before library preparation, the quality of RNA samples was assessed with an Agilent Bioanalyzer RNA 6000 pico chip, and the RNA concentrations were measured with a Qubit fluorometer. RNA library preparation was performed using Illumina Stranded mRNA Prep with ribosomal RNA depletion and PolyA purification. The libraries were sequenced on Illumina NextSeq 2000 to generate paired-end reads of 150x2 bases. The raw data were demultiplexed using bcl-convert 4.1.5, and then the adapters were trimmed using cutadapt 3.2 (M. Martin, 2011).

3. Transcriptome *de novo* assembly and annotation

We preprocessed the newly obtained raw reads (root and nodule *I. argentea*) and publicly available SRR datasets (nodules of *Lupinus mariae-josephae* and *L. luteus*) to ensure high-quality data for the downstream analysis. Briefly, the remaining adaptors were removed with Fastp version 0.20.0 (Chen, 2023), reads with unfixable errors removed with Rcorrector version 1.0.4 (Song & Florea, 2015) , and FilterUncorrectablePEfastq.py python script (github.com/harvardinformatics/), and the remaining short or low-quality reads (Q score < 20 and length < 25) removed with TrimGalore version 0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) We assessed the read quality after each preprocessing step with FastQC (Andrews, 2010). To remove possible bacterial contamination from *I. argentea* root and nodule read sets, we mapped our reads using Bowtie2 version 2.3.5.1 (Langmead & Salzberg, 2012) against the *Bradyrhizobium elkanii* SA281 bacterial strain that we used to inoculate our plants. We then used Samtools version 1.10 (Danecek et al., 2021) to keep only unmapped reads and convert our data to Fastq format.

The transcriptomes of all four RNAseq datasets were de novo assembled separately following the same process. First, we provided our processed clean reads of the three replicates simultaneously to Trinity version 2.6.6 (Grabherr et al., 2011) to generate de novo transcriptome assembly. Second, in order to merge the *Indigofera argentea* roots and nodules contigs, we used CD-HIT version 4.8.1 (Fu et al., 2012). Then, we clustered our contigs into gene-level clusters, first using Corset version 1.0.9 (Davidson & Oshlack, 2014) to find gene isoforms and then Lace version 1.14.1 (Davidson et al., 2017) to merge those isoforms to form super transcripts.

We checked the quality of the Trinity and Supertranscripts assemblies using STAR version 2.7.3a (Dobin et al., 2013) to obtain the alignment rate and BUSCO version 5.5.0 (Simão et al., 2015) to assess the completeness of the assemblies using the 'Viridiplantae' and 'Fabales' databases. For *I. argentea*, we checked the quality of the assemblies (alignment rate and completeness of the assemblies) before and after merging them to check the quality of the merged assembly.

We predicted the coding regions using the TransDecoder version 5.7.1 (https://github.com/TransDecoder/TransDecoder) based on the results of the Blastp version 2.12.0+ (Camacho et al., 2009) search against the UniProt database (release-2023_05) supplemented with all the known NCR peptides (Czernic et al., 2015; Montiel et al., 2017; Raul et al., 2021). Finally, we performed the functional annotation of *Indigofera argentea* protein sequences using the GFAP (Gene Functional Annotation for Plants) pipeline using *Glycine max* species (the closest reference plant species available in the database) (D. Xu et al., 2023). The AHL and LegHB proteins were annotated manually using tBlastn and all AHL and LegHB from legumes from NCBI.

4. Homology, orthology, and clustering analysis

The genomic and predicted proteomes data of legume species with known NCR peptides (four IRLC and two Dalbergioid species) were collected from NCBI and the Legume Information System (LIS, https://www.legumeinfo.org/) (Berendzen et al., 2021). The known IRLC and Dalbergioid NCR peptide sequences were collected from (Montiel et al., 2017) and (Czernic et al., 2015; Raul et al., 2021), respectively.

To ensure the presence of NCR peptides in the predicted proteomes, we searched for the known NCR peptides in their corresponding species predicted proteome using blastp (Camacho et al., 2009). If the NCR peptides are not found on the proteome, tBlastn was used to search for them in their corresponding species genome. We added them to the proteome if they were present in the genome.

Using the predicted proteomes, including all known NCR peptides, we scored the similarity among all sequences of all species using blastp (Camacho et al., 2009) and used those scores to define a set of orthologs using orthAgogue software (Ekseth et al., 2014) and then regrouped those orthologs into clusters using Markov Clustering (MCL) (L. Li et al., 2003). Custom scripts were used to extract NCR clusters from all protein clusters. We defined an "NCR cluster", each

cluster that has at least two NCR peptides. In the NCR-mixed clusters with at least two NCRs and other proteins, we extracted only NCR peptides.

5. NCR peptide detection and classification

The DNA and protein sequences of NCR clusters were extracted using custom scripts. SignalP version 4.0 (Petersen et al., 2011) was used to exclude clusters where there is no signal peptide in at least two NCR peptides. Macse2 software was used to produce codon-based multiple sequence alignments from CDS sequences of the remaining NCR clusters. Translated protein Hidden Markov Models (HMM) profiles were built from those alignments with hmmbuild from hmmer version 3.3.2 (Johnson et al., 2010).

The legume species used to search NCR peptides were the above *de novo* assembled transcriptomes (*Indigofera argentea*, *Lupinus luteus*, and *Lupinus mariae-josephae*), the assembled nodule transcriptomes from (Huang et al., 2022) (*Medicago sativa* and *Melilotus officinalis*) and from (Kant et al., 2016) (*Cicer arietinum),* and the publicly available legume genomes and nodule RNA-sequencing data (*Medicago truncatula*, *Pisum sativum*, *Trifolium pratense*, *Arachis hypogaea, Aeschynomene evenia, Lupinus albus, Lotus japonicus, Cajanus cajan, Phaseolus vulgaris, Vigna angularis,* and *Glycine max*).

The RNA-seq data were downloaded from the Sequence Read Archive (SRA) database. They were converted to fastq files using the fastq-dump tool from the SRA Toolkit. Their corresponding genomes were also downloaded from NCBI and LIS.

As described above, NCR peptides are small peptides, and their sequences are highly divergent. Therefore, known functional annotation methods are not sensitive enough to detect NCR peptides from genomes or transcriptomes. We therefore used SPADA (Small Peptide Alignment Discovery Application) pipeline version 1.0 (P. Zhou et al., 2013) to search for NCR peptides in our genomes and nodules transcriptomes. SPADA is a computational pipeline that, when provided with multiple sequence alignments for a particular gene family, identifies all members of this family in a target genome sequence. SPADA pipeline is specialized in predicting cysteine-rich peptides in plant genomes. First, we used the "seq.check" command of SPADA to check our genomes. Second, we used the command "build_profile" to build HMM profiles from our IRLC and Dalbergioid NCR cluster alignments. We then ran SPADA three times separately for each genome or assembled nodule transcriptome, one using IRLC profiles, one using Dalbergioid profiles, and one using CRP profiles from the SPADA pipeline, which

are plant cysteine-rich peptides that we used because we were not sure that our clade-specific clusters could capture all NCRs in other clades. For each genome, we merged the results from the three analyses, converted the merged results to fasta format, and removed the duplicates.

The predicted putative NCR peptides were then filtered according to their length, cysteine motif, and their expression in nodules. To predict the length of mature peptides, we used signalP version 4.1 (Petersen et al., 2011) with the "notm" network to predict the cleavage sites and extract the mature peptides. Using custom scripts, the length of the mature peptides and the number of cysteines were counted, and we kept only NCR peptides whose mature peptides have fewer than 100 aa and at least four cysteines. We then assessed the expression in nodules by mapping the reads of nodule RNA-seq data against the genome using STAR version 2.7.3a (Dobin et al., 2013) and quantifying the number of reads using htseq-count version 0.12.3 (Putri et al., 2022). We kept only putative NCR peptides expressed in nodules.

We classified the retained NCR peptides into two groups: NCRs with four or six cysteines (NCR-motif) and NCRs with eight cysteines (defensin-motif). For the IRLC species, we annotated as NCR only those with the NCR motif, as defensin motif-containing peptides are presumed to act in the innate immune system of IRLC plants. However, for the other clades, we annotated both NCR-motif and defensin-motif as NCR. To compute the pI values of the NCR peptides and clusters, we used the R package pIR, where the pI of each peptide was calculated based on the mean values from all prediction methods, excluding the highest and lowest values (Audain et al., 2016).

We then performed a classification step for NCRs found with CRP (Cysteine-Rich Peptide) profiles (SPADA did not classify them with our IRLC or Dalbergioid profiles), where we searched them against our cluster profiles using hmmsearch version 3.3.2 (S. Eddy, 2009) and chose the best hit profile to classify each sequence. Finally, we merged NCR peptides classified with SPADA and those classified with the HMMsearch approach for each species.

Moreover, for *I. argentea,* where only a few NCR peptides were found, and all of them were unclassified, we ran SPADA a second time using our expanded clusters, including newly found NCRs in IRLC, Dalbergioids and Genistoids and also using one NCR HMM profile of *I. argentea* that we built using the five NCR peptides found in (Ren, 2018) The filtration steps here were more flexible where we did not exclude long NCR peptides that are differentially expressed in nodules and have an NCR cysteine motif.

6. Differential expression analysis

In order to estimate the gene expression levels of *I. argentea* and to highlight the differentially expressed NCR peptides between root and nodules, we used the mapping-based mode of salmon version 0.12.0 (Patro et al., 2017) to map our samples individually to the reference transcriptome described above. We used the R package DESEQ2 version 1.32.0 (Love et al., 2014) to predict the differentially expressed genes between root and nodules of *I. argentea* using the raw counts for each replicate found with salmon, the length of each gene for normalization, and the annotation of the genes. We also used the diCoexpress R package (Lambert et al., 2020), which performs differential expression analysis using generalized linear models with the edgeR package (M. D. Robinson et al., 2010). The genes with log2(FoldChange) > 2 and the adjusted p-value < 0.05 were considered up-regulated, and the genes with log2(FoldChange) < 0.5 and the adjusted p-value < 0.05 were considered down-regulated.

7. Structural and phylogenetic analysis

A full command-line local installation of Alphafold2 version 2.2 (Jumper et al., 2021) on the I2BC server was used to predict the 3D structure of the classified NCR peptides (one per cluster). We predicted the 3D structures of one randomly picked NCR peptide per cluster. Structural models were constructed using Alphafold2's own embeddings, supplemented with macse-generated multiple sequence alignments of sequence-based clusters. All templates downloaded on June 30, 2022, were allowed for structure modeling. Five models were predicted, and we selected the model with the best pLDDT (predicted Local Distance Difference Test) score. For further analysis, we kept only the NCR structures with a pLDDT score higher than 70. For the non-classified predicted NCR peptides from Indigoferoids (*Indigofera argentea)* and Genistoids (*Lupinus* spp), we first used the CD-HIT version 4.8.1 (Fu et al., 2012) with an identity threshold of 90, 80, 70, and 50% to regroup them into clusters. However, all the non-classified NCRs were monotypic, sharing no similarities with each other. Thus, we predicted the 3D structure of each of them using AlphaFold2, as described above, and we included them for further analysis in our dataset of the NCR clusters as monotypic NCRs. In order to check if NCR peptides evolved from defensins or not, we predicted the 3D structures of 48 defensins from at least one species per clade, and we included them in our structural comparison analysis.

We then used Foldseek version 4-645b789 (Barrio-Hernandez et al., 2023), a structure clustering approach to regroup all the high-quality NCR and defensin structures into superclusters. Foldseek is a new and fast method that converts the 3D structures into 1D vectors that contain the structure information, which decreases computation times by four to five orders of magnitude. The TM scores within each supercluster and between each supercluster and all others were also computed with foldseek to compare intra-supercluster versus inter-supercluster TM scores, and the distributions of TM scores were plotted in R with ggplot. For each supercluster, a structural alignment of all the structures was computed using USalign (C. Zhang et al., 2022) or MUSTANG (Konagurthu et al., 2006). These structural alignments were represented with a sausage representation with a blue-red color gradient that represents the TM score ranging from 50 to 100 of the all-vs-all alignment of all the structures of the superclusters, but represented by the alignment of all structures to the longest structure (the TM scores are the scores from all-vs-all alignment, but for a better visualization we kept only the alignment of all structures to the longest structure).

Then, in order to generate a structural "phylogenetic" tree based on the structural distances of all the NCR and defensin structures, and for each supercluster, we used Foldtree (Moi et al., 2023). Foldtree uses the all-vs-all comparison of Foldseek and creates a distance matrix from the Fident scores (sequence similarity after aligning with the structural alphabet) of Foldseek output, which is used as the input of quicktree (Howe et al., 2002) to generate the structure-based tree. The Foldseek parameters used for the Foldtree are 0.5 of coverage, 0.25 of the Foldseek alphabet sequence identity, an e-value of 0.1, and an exhaustive search. These parameters were chosen based on the separation of the superclusters and the congruence of the tree after testing more than 100 combinations of parameters. Furthermore, in order to avoid any noise in our analysis and to have congruent trees, we excluded the monotypic superclusters. The trees were analyzed and annotated with iTol version 5 (Letunic & Bork, 2021).

A sequence similarity network of the representative sequence of each cluster (the same predicted by Alphafold2 used for the structural analysis) was also produced using CLANS (CLuster ANalysis of Sequences) (Frickey & Lupas, 2004). The network was annotated manually with the superclusters, and the same annotation was used in the structural phylogenetic tree.

8. Flow cytometry

Bacteroid extraction was performed as described before in (Mergaert et al., 2006). Bacteroids and free-living *Bradyrhizobium elkanii* SA281 bacteria were strained with 50 µg/ml DAPI in BEB. After 10 minutes of incubation at room temperature, the bacteria and bacteroids were processed with a Cytoflex cytometer (Beckman-Coulter). The data analysis was performed using cytExpert software v2.5 and the flowCore package in R (https://bioconductor.org/packages/release/bioc/html/flowCore.html).

9. Confocal microscopy

Nodules live-dead imaging was performed on an SP8X confocal DMI 6000 CS inverted microscope (Leica). First, fresh nodules were harvested, embedded in 6% agarose, and sliced into 70µm slices using a Leica vibratome. The slices were incubated for 15 minutes in a 50mM Tris-HCl buffer with 0.01% calcofluor white (to strain plant cells) M2R (Sigma), containing 0.5 µl of Syto9 (to stain live bacteria) and 0.5 µl of Propidium iodide (to stain dead bacteria). Then, the washed sections were observed with ×10 dry and ×63 oil immersion objectives. The analysis of images was performed using ImageJ software (T. J. Collins, 2007).

10. *In vitro* NCR sensitivity assays

All chemically synthesized peptides were purchased from Genscript. The purity of all peptides is > 98%. The heme assays were performed as described in (Sankari et al., 2022). The growth assays of *S. meliloti* and *E. coli* were also described in (Sankari et al., 2022). Briefly, overnight cultures were washed and diluted in GSY media. The diluted cultures were distributed in sterile 96 well plates, and OD_{600nm} was measured every hour using a Tecan SPARK 10M microplate reader with continuous shake at 150 rpm. To check the effect of NCR peptides on the cell cycle (i.e. genome amplification), we quantified the DNA content of *S. meliloti* supplemented with each NCR peptide using flow cytometry as described previously in (Haag et al., 2011).

ACKNOWLEDGEMENTS

A.B. benefited from a Ph.D. contract in the frame of the CNRS 80|PRIME – 2021 program (awarded to T.G. and P.M.) and was partially supported by a Mitacs Globalink Research Award. B.A. benefited from a French State grant (Saclay Plant Sciences, reference n° ANR-17-EUR-

0007, EUR SPS-GSR) under a France 2030 program (reference n° ANR-11-IDEX-0003). We thank Rui Huang for sharing his scripts that were used with slight modifications.

2. Taxonomic distribution of SbmA/BacA and BacA-like antimicrobial peptide transporters suggests independent recruitment and convergent evolution in host-microbe interactions

Authors: Nicholas T. Smith[1]+, Amira Boukherissa[2,3]+, Kiera Antaya[1], Graeme W. Howe[4], Ricardo C Rodríguez de la Vega[3], Jacqui A. Shykoff[3], Benoît Alunni[2,5]*, George C. diCenzo[1]*

+ N.T.S. and A.B. contributed equally to this work.

Affiliations:

1 Department of Biology, Queen's University, Kingston, ON, K7L 3N6, Canada

2 Institute for Integrative Biology of the Cell, CNRS, CEA, Université Paris-Saclay, 91198, Gif-sur-Yvette, France

3 Écologie Systématique et Évolution, CNRS, Université Paris-Saclay, AgroParisTech, 91198, Gif-sur-Yvette, France

4 Department of Chemistry, Queen's University, Kingston, ON, K7L 3N6, Canada

5 Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France

*** Corresponding authors:** Benoît Alunni (benoit.alunni@inrae.fr) and George C. diCenzo (george.dicenzo@queensu.ca)

Keywords: SbmA, antimicrobial peptides, host-microbe interaction, peptide transport, convergent evolution, rhizobium-legume symbioses, pathogenesis

A. Foreword

This work also started at the beginning of my thesis. It resulted from a collaboration with Dr. George diCenzo and Nicolas Smith from the diCenzo lab at Queen's University in Canada.

In this project, we studied the taxonomic distribution of BacA and BclA transporters across the bacterial domain and their repeatable evolution for NCR resistance in rhizobium-legume symbiosis. We know that these two transporters, along with the YejABEF transporters, are required for TBD in NCR-triggered legume-rhizobium symbiosis. We also know that BacA and BclA transporters are phylogenetically distant and have different mechanisms of transport due to their structural differences. However, despite this difference, both proteins can import NCR peptides promoting TBD. So here, we studied those transporters' evolution and taxonomic distribution to answer whether bacterial BacA and BclA proteins followed the same evolutionary path for NCR resistance in rhizobia.

During this work, we combined molecular biology experiments, in-planta assays, and bioinformatic analysis to gain insights into the function and evolution of these peptide transporters. First, I collected the genomic and proteomics data of bacteria. I used statistical and phylogenetic analysis to extract BacA transporters and infer the evolutionary history of BacA transporters. I also used the Multi Locus Sequence Analysis (MLSA) to infer the evolution of the bacteria and study the taxonomic distribution of those transporters across the bacterial domain. Second, we selected nine distant *bclA* genes and one *bacA* to test their function *in vitro.* Nicholas Smith cloned those nine genes into the pRF771 expression vector. He also transferred those plasmids into *S. meliloti* Δ*bacA* mutant and performed gentamicin sensitivity assays. Then, in order to check their ability to import NCR peptides, I transferred those plasmids into *S. meliloti* Δ*bacA* Δ*yejA* and performed NCR247 sensitivity assays using both the simple and double-complemented mutants. The function of those genes was also tested *in planta* by Nicholas Smith using the alfalfa and sweet clover legumes inoculated by the S. *meliloti* Δ*bacA* complemented mutants. The paper has been deposited in Microbial Genomics journal (N. T. Smith et al., 2024).

B. Abstract

Antimicrobial peptides (AMPs) are often produced by eukaryotes to control bacterial populations in both pathogenic and mutualistic symbioses. Several pathogens and nitrogen-fixing legume symbionts depend on transporters called SbmA (or BacA) or BclA (BacA-like) to survive exposure to AMPs. However, how broadly these transporters are distributed amongst bacteria, and their evolutionary history, is poorly understood. We used computational approaches to examine the distribution of SbmA/BacA and BclA proteins across 1,255 bacterial species, leading to the identification of 71 and 177 SbmA/BacA and BclA proteins, respectively. Phylogenetic and sequence similarity analyses suggest that the functional similarity of the SbmA/BacA and BclA protein families is likely due to convergent evolution. In vitro sensitivity assays using a legume AMP and several of the BclA proteins confirmed that AMP transport is a common feature of BclA proteins. Our analyses indicated that SbmA/BacA orthologs are encoded only by species in the phylum Pseudomonadota and are primarily found in just two orders: Hyphomicrobiales and Enterobacterales. BclA orthologs are somewhat more broadly distributed and were found in clusters across four phyla. These included several orders of the phyla Pseudomonadota and Cyanobacteriota, as well as the order Mycobacteriales (phylum Actinomycetota) and the class Negativicutes (phylum Bacillota). Many of the clades enriched for species encoding Sbma/BacA or BclA orthologs are rich in species that interact with eukaryotic hosts in mutualistic or pathogenic interactions. These observations suggest that SbmA/BacA and BclA proteins have been repeatedly co-opted to facilitate associations with eukaryotic hosts by allowing bacteria to cope with host-encoded AMPs.

C. Introduction

Bacteria, whether they thrive as free-living organisms or in interaction with eukaryotic hosts, are constantly challenged with a variety of stresses including exposure to antimicrobial peptides (AMPs). These peptides are usually ca. 10-60 aa, and although they vary in their amino acid compositions, AMPs of the same family often display an enrichment in a specific amino acid such as cysteine (forming intramolecular disulfide bridges), proline, arginine, or glycine. These peptides may have two main modes of actions depending on their cellular targets. Membrane-damaging AMPs interact with lipid bilayers and insert into biological membranes, thereby forming pores leading to cell content leakage and loss of ion gradients and membrane potential (Brogden 2005). These are mostly cationic peptides, whose charge is involved in establishing an interaction with biological membranes resulting in their destabilization. Other AMPs have intracellular targets and disturb bacterial metabolism and physiology, notably by interacting with metabolic enzymes, transcriptional and translational machineries, and cell cycle regulators (Le et al. 2017). Bacteria have evolved several mechanisms to cope with AMPs, including the expression of dedicated antimicrobial peptide transporters (Gebhard 2012; Gruenheid and Le Moual 2012).

An example of a bacterial transporter able to import AMPs is SbmA, which was originally identified in *Escherichia coli* as conferring resistance to Microcin B17 (Laviña et al. 1986; Salomón and Farías 1995). Interestingly, homologs of SbmA (often known as BacA), and the related protein BclA (standing for BacA-like), have been identified in other bacteria interacting with eukaryotic hosts such as plant symbionts (eg. nitrogen-fixing rhizobia) and animal/human pathogens (*Brucella abortus*, *Mycobacterium tuberculosis*) (Glazebrook et al. 1993; LeVier et al. 2000; Domenech et al. 2009). SbmA/BacA and BclA are inner membrane peptide transporters, differing primarily by the presence of an ATPase domain in BclA that is absent in SbmA/BacA (Guefrachi et al. 2015). ATP hydrolysis by the ATPase domain is essential for the transport activity of BclA, whereas BacA-mediated transport is driven by the proton-motive force (Runti et al. 2013; Ghilarov et al. 2021). SbmA/BacA and BclA can import (and possibly export (Nijland et al. 2024) a variety of AMPs, including nodule-specific cysteine-rich (NCR) peptides produced by some legume plants (Haag et al. 2011) and proline-rich mammalian peptides (Mattiuzzo et al. 2007). They can also transport several non-proteinaceous compounds like the antibiotic gentamicin (LeVier and Walker 2001) and the vitamin cobalamin (Nijland et al. 2024). Another AMP transporter is the bacterial YejABEF transporter. Like SbmA/BacA and BclA, YejABEF can import a range of AMPs and has been found in both plant mutualists

and eukaryotic pathogens (Novikova et al. 2007; Eswarappa et al. 2008; Wang et al. 2016; Nicoud et al. 2021).

As many AMP transporters can import a range of AMPs and considering the differing mechanisms of actions of diverse AMPs, these transporters have positive or negative effects on fitness depending on the environment. A good example is the role of BacA in *Sinorhizobium meliloti*, which is a nitrogen-fixing symbiont of legumes like *Medicago truncatula*. During symbiosis with legumes, *S. meliloti* cells reside intracellularly within a specialized legume structure called a root nodule. Some legumes, like *M. truncatula*, produce a family of cysteine-rich AMPs known as NCR peptides, whose isoelectric points (pI) vary from 3 (anionic) to 11 (cationic) (Van De Velde et al. 2010; Czernic et al. 2015; Montiel et al. 2017; Kereszt et al. 2018; Huang et al. 2022). *S. meliloti* strains carrying loss-of-function *bacA* mutations are hypersensitive to cationic NCR peptide exposure *in vitro* (Haag et al. 2011) and die rapidly upon release into *M. truncatula* nodules in an NCR peptide-dependent fashion (Glazebrook et al. 1993; Haag et al. 2011). It has been hypothesized that by importing NCR peptides, BacA moves the cationic NCR peptides away from the cell membrane, thereby protecting *S. meliloti* from the membrane-damaging activities of these AMPs and promoting fitness (Haag et al. 2011; Farkas et al. 2014; diCenzo et al. 2017). Similarly, BacA/BclA homologs are required by *B. abortus*, *M. tuberculosis*, and *E. coli* for chronic infection of their eukaryotic hosts (LeVier et al. 2000; Li et al. 2005), likely in a similar fashion. On the other hand, phazolicin is an AMP produced by the bacterium *Rhizobium* sp. Pop5, which is toxic to *Rhizobium* and *Sinorhizobium* strains due to its ability to inhibit translation intracellularly (Travin et al. 2019). The import of phazolicin by *S. meliloti* is mediated by the BacA and YejABEF transporters, with mutation of both transporters resulting in resistance to this AMP (Travin et al. 2023). Thus, there is likely a fitness trade-off to *S. meliloti* encoding BacA; its presence increases fitness during legume symbiosis but may decrease fitness in the soil in the presence of AMPs produced by other microbes. More broadly, a recent study predicted nearly one million new AMPs from microbiome data (Santos-Júnior et al. 2024), suggesting that bacteria encounter many diverse AMPs in environmental niches. It is therefore likely generally true that encoding AMP transporters like SbmA/BacA come with a fitness trade-off, where these transporters promote fitness in the presence of membrane-targeting AMPs but impair fitness in the presence of AMPs with intracellular targets.

The observation that SbmA/BacA and BclA orthologs are found in diverse bacterial lineages suggests that these proteins may be widespread housekeeping proteins subsequently co-opted

for host-bacterial interactions (Arnold et al. 2013). On the other hand, the potential fitness trade-offs means that the maintenance of these genes likely depends on the types of AMPs that a given bacterium encounters. However, no systematic study of the distribution of SbmA/BacA or BclA orthologs across the bacterial tree exists. In addition, the evolutionary relationship between the SbmA/BacA and BclA families remains to be elucidated. Here, we report the distribution of SbmA/BacA and BclA orthologs in 1,255 bacterial species from across the bacterial domain. We found SbmA/BacA orthologs exclusively within the phylum *Pseudomonadales* (syn. *Proteobacteria*), while BclA orthologs were predominately limited to the phyla *Pseudomonadales*, *Cyanobacteriota* (syn. *Cyanobacteria*), *Actinomycetota* (syn. *Actinomycetes*), and *Bacillota* (syn. *Firmicutes*). Expression of a subset of the newly identified BclA proteins in *S. meliloti* Δ*bacA* mutants confirmed that transport of antimicrobial peptides is a common property of the BclA protein family. The taxonomic distribution of SbmA/BacA and BclA, together with phylogenetic analysis of these proteins, leads us to suggest that the functional similarities between SbmA/BacA and BclA are a result of convergent evolution, and that these protein families have been repeatedly co-opted to help microbes cope with antimicrobial peptide exposure during host-microbe interactions and possibly in prokaryote-prokaryote interactions.

D. Results

1. Identification and classification of SbmA/BacA and BclA orthologs across the bacterial domain

To study the evolution and distribution of SbmA/BacA and BclA proteins, we searched the proteomes of 1,255 bacterial species, each belonging to a distinct genus, for proteins showing similarity to the SbmA/BacA-like family of PFAM (PF05992) (see Materials and Methods). This process led to the identification of 366 putative SbmA/BacA-like family proteins from 258 species. We further classified each of these 366 proteins into one of five protein classes according to Guefrachi and colleagues (Guefrachi et al., 2015): SbmA/BacA, BclA, *Mycobacterium* BacA (a BclA-like family of proteins first identified in *M. tuberculosis*), ExsE (a related protein family involved in long-chain fatty acid transport), and the so-called *Bradyrhizobium* homologous clade (a related protein family with an unknown function). Initially, this classification was based on the use of hidden Markov models (HMMs), which was subsequently refined based on phylogenetic reconstruction and a sequence similarity network (SSN) as described below.

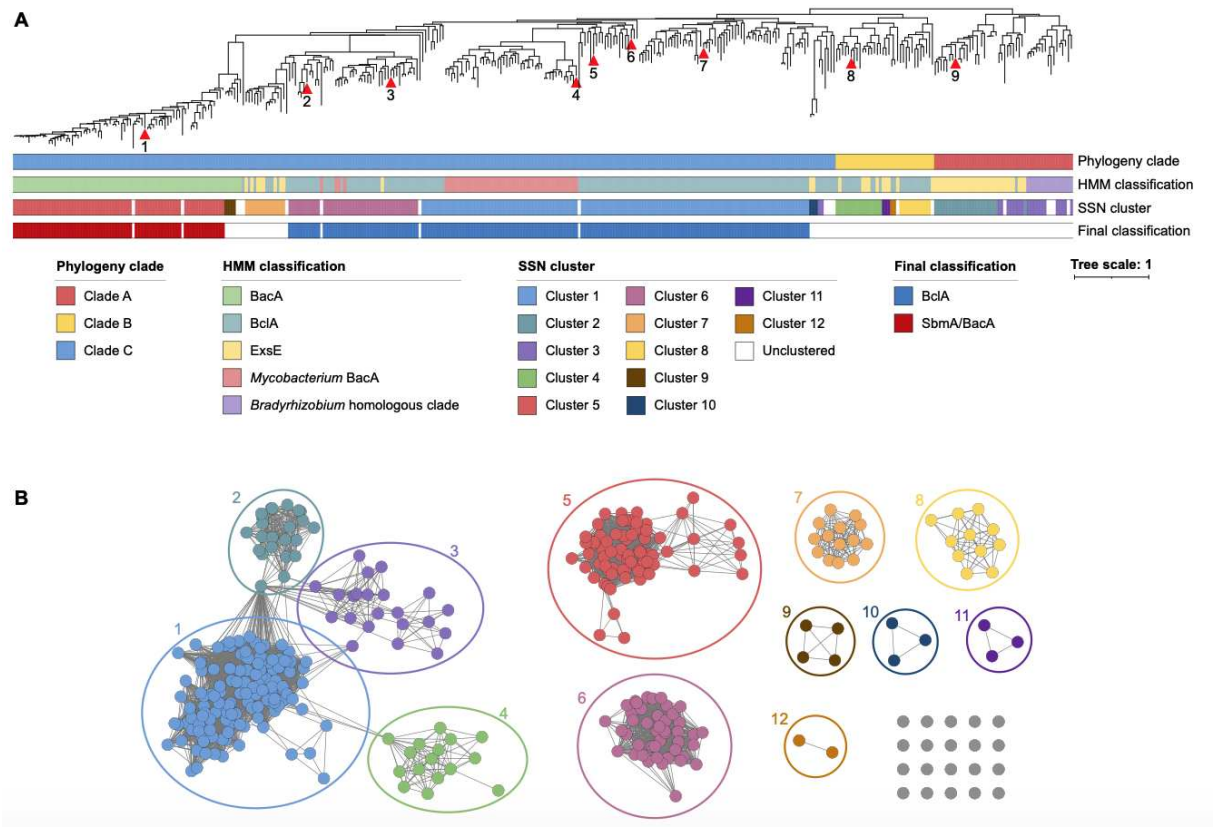


Figure 33 Sequence and phylogenetic analysis of SbmA/BacA-like proteins.

(**A**) An unrooted maximum likelihood phylogeny of 366 SbmA/BacA-like proteins is shown. The scale bar represents the average number of amino acid substitutions per site. Red triangles indicate proteins whose corresponding genes were codon optimized and synthesized: 1 - *Polymorphum gilvum* BclA; 2 - *Synechococcus elongatus* BclA; 3 - *Cyanobacterium aponinum* BclA; 4 - *Basilea psittacipulmonis* BclA; 5 - *Succinivibrio dextrinosolvens* BclA; 6 - *Methylomusa anaerophila* BclA; 7 - *Polaromonas naphthalenivorans* BclA; 8 - *Eikenella exigua* BclA-like; 9 - *Phyllobacterium zundukense* ExsE. The bars beneath the phylogeny summarize the clustering and annotation of these proteins. The top bar indicates the phylogenetic clade to which each protein belongs. The second bar indicates the preliminary hidden Markov model (HMM) classification of each protein. The third bar indicates the cluster in the sequence similarity network that each protein belongs to. The bottom bar indicates which proteins were ultimately classified as SbmA/BacA (red) or BclA (blue). An interactive version of this phylogeny, with node support values, is provided through iTOL (https://itol.embl.de/shared/1IAjjFrHYGLI9) while a Newick-formatted version of the phylogeny can be downloaded from GitHub (https://github.com/amira-boukh/SbmA_BacA_phylogenetic_distribution). (**B**) A sequence similarity network calculated using EFI-EST of 366 SbmA-BacA-like proteins is shown. Each node (the circles) represents one protein, while edges (the lines) represent sequence similarity between pairs of proteins above the threshold, with longer lines indicating lower similarity. Nodes are color-coded based on cluster.

Using HMMs for these five protein classes, the 366 SbmA/BacA-like family proteins were initially classified into 79 SbmA/BacA proteins, 169 BclA proteins, 50 *Mycobacterium* BacA proteins, 52 ExsE proteins, and 16 *Bradyrhizobium* homologous clade proteins (**Figure 33A**).

A maximum-likelihood phylogenetic analysis led to the identification of three primary monophyletic groups (**Figure 33A**). Clade A comprised 48 proteins and included most ExsE and all *Bradyrhizobium* homologous clade proteins, which we treated as the outgroup. Clade B included 34 proteins that were annotated as a mix of BclA and ExsE based on the HMMs. Clade C was the largest clade, consisting of 284 proteins, and included most of the putative BclA, SbmA/BacA, and *Mycobacterium* BacA proteins.

Most of the putative BclA proteins from Clade C also form a single cluster in the SSN (Cluster 1; **Figure 33B**). We, therefore, conclude that the 133 proteins of Cluster 1 in the SSN represent true BclA orthologs. Notably, Cluster 1 of the SSN also includes 46 proteins annotated as *Mycobacterium* BacA, which also fall within Clade C in the phylogeny (**Figure 33**). This suggests that the *Mycobacterium* BacA proteins are not a distinct family from the BclA proteins and that *Mycobacterium* BacA proteins should instead be referred to as BclA. On the other hand, a Clade C subclade of nine proteins with long branch lengths in the phylogeny is excluded from Cluster 1 of the SSN; instead, two of these proteins are found as part of Cluster 3 that, predominantly consists of the *Bradyrhizobium* homologous clade proteins, three are found as a three-protein cluster (Cluster 10), and five are singletons. In addition, four of these nine proteins are from strains encoding a BclA protein belonging to Cluster 1. Taken together, we conclude that these nine proteins are not true BclA orthologs. Another subclade of Clade C consisting of 58 proteins is not part of Cluster 1 in the SSN but rather is largely found in two clusters (Clusters 6 and 7) of 44 and 14 proteins, respectively (**Figure 33**). Cluster 6 consists primarily of proteins from cyanobacteria, and 43 of the 44 proteins were classified as BclA or *Mycobacterium* BacA by the HMMs. In addition, the functional data described below suggests that proteins of this cluster are functionally similar to known BclA proteins. We, therefore, conclude that proteins of Cluster 6 represent BclA orthologs. In contrast, eight of the 14 proteins of Cluster 7 were annotated as ExsE by the HMMs. The distinct clustering of Cluster 7 from Cluster 6, together with the HMM annotations, leads us to suggest that the proteins of Cluster 5 are unlikely to represent true BclA orthologs.

Consistent with the phylogenetic analysis, proteins of Clade B do not cluster with proteins of Clade C in the SSN (**Figure 33B**). Rather, the Clade B proteins are split across four clusters and two singletons. Nearly 1/3rd (10 of 34) proteins of Clade B were annotated as ExsE by the initial HMM strategy, and many of the proteins of Clade B are from bacterial strains that also encode a putative SbmA/BacA or BclA of Clade C. Collectively, we interpret these results to indicate that Clade B proteins are not part of the BclA protein family and that they instead

represent a related but distinct protein family. This conclusion is also supported by the functional data presented below.

Lastly, all putative SbmA/BacA proteins formed a monophyletic group in the phylogeny (**Figure 33A**), and a monophyletic group of 71 of the 79 proteins formed a single cluster (Cluster 5) in the SSN (**Figure 33B**). These results suggest that the 71 proteins of Cluster 5 and annotated as SbmA/BacA by the HMM strategy are likely true SbmA/BacA orthologs and that all SbmA/BacA proteins evolved from a common ancestor. Although the SbmA/BacA proteins fell within Clade C in the phylogeny, the SbmA/BacA clade is connected to the rest of the tree via an unusually long branch, consistent with the distinct clustering of SbmA/BacA proteins in the SSN. The distinct clustering in the SSN, the long branch length, and the functional differences in transport (ATP-driven vs proton-driven) lead us to suggest that the SbmA/BacA and BclA protein families evolved independently and that their functional similarity is a result of convergent evolution.

In considering the different sources of information described above, we ultimately chose to select a final set of SbmA/BacA and BclA proteins based primarily on the SSN, resulting in the identification of 177 high-confidence BclA proteins (including the *Mycobacterium* BacA proteins) and 71 high-confidence BacA proteins (**Figure 33A**).

2. In vitro functional analysis of diverse SbmA/BacA and BclA orthologs

To validate that the BclA and SbmA/BacA proteins identified through the *in silico* approach are functionally similar to known BclA and SbmA/BacA proteins, genes encoding nine of the identified proteins were synthesized. The proteins encoded by these genes included one BacA protein, six BclA proteins, including one previously classified as *Mycobacterium* BacA, one protein from Clade B (henceforth referred to as BacA-like), and one ExsE protein for comparison. The nine genes were then cloned into an expression vector and introduced into *S. meliloti* Δ*bacA* and *S. meliloti* Δ*bacA* Ω*yejA* mutants to test for complementation. Although the genes were codon-optimized for expression in *S. meliloti*, we cannot exclude the possibility that some proteins were not properly expressed or were not stably inserted into the *S. meliloti* inner membrane. Therefore, a lack of complementation may reflect improper expression/localization of a protein rather than a lack of orthology. All strains showed similar growth in media lacking antimicrobial agents (**Figure S4**), indicating that differences in media supplemented with gentamicin (Gm) or NCR peptides reflect altered resistance phenotypes rather than general growth differences.

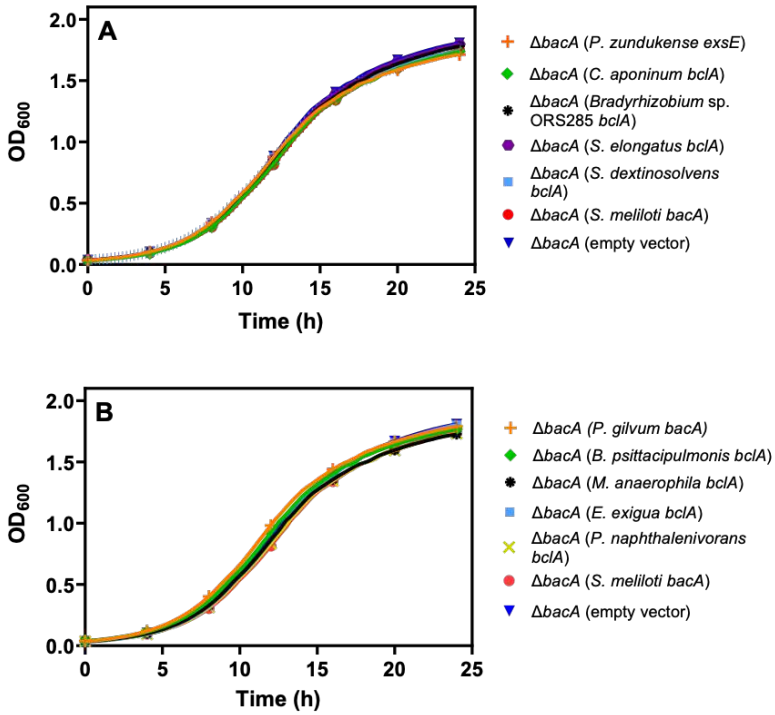


Figure S 4 Growth of Sinorhizobium meliloti strains in LBmc.

The growth of various *S. meliloti* strains, as measured by OD600, in LBmc is shown over a 24-hour period. Each point represents the mean of triplicate wells, with error bars depicting standard deviation. The Δ*bacA* strain represents the *S. meliloti* Δ*bacA* mutant carrying an empty vector, while all other strains are named according to the species of origin of the gene expressed *in trans* in the *S. meliloti* Δ*bacA* background. The experiment was replicated three independent times, and data from a representative experiment is shown.

In addition, we observed that the resistance phenotypes of the *S. meliloti* Δ*bacA* mutant complemented with the *S. meliloti bacA* gene *in trans* differed somewhat from wildtype *S. meliloti* (**Figure S5**), likely due to elevated expression of *bacA* in the complemented strain. Thus, for all *in vitro* phenotypic experiments, strains were compared to the *S. meliloti* Δ*bacA* mutant complemented with the *S. meliloti bacA* gene *in trans* rather than the wild type.

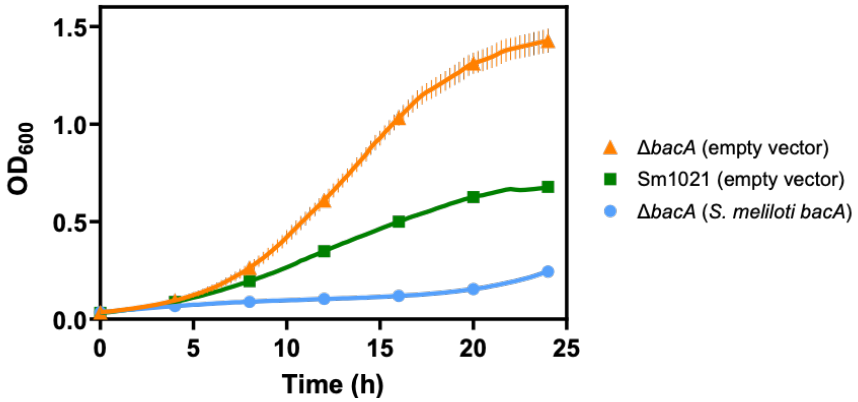


Figure S 5 Effect of expressing bacA in trans on the gentamicin sensitivity of Sinorhizobium meliloti

The growth of three *S. meliloti* strains, as measured by OD600, in the presence of 20 µg/mL of gentamicin is shown over a 24-hour period. Growth profiles are shown for wildtype *meliloti* Sm1021 harboring an empty expression vector (blue), and a *S. meliloti* Δ*bacA* mutant expressing (blue) or not (orange) the *S. meliloti bacA* gene *in trans*. Each point represents the mean of triplicate wells, with error bars depicting standard deviation. The experiment was replicated three independent times, and data from a representative experiment is shown.

As *S. meliloti bacA* null mutants display increased resistance to Gm (LeVier and Walker 2001), we first tested whether the nine genes could complement the Gm resistance phenotype of the *S. meliloti* Δ*bacA* mutant. As expected, the Δ*bacA* mutant was resistant to Gm, and reintroduction of the *S. meliloti bacA* gene *in trans* resulted in sensitivity to Gm (**Figure 34A**). Unexpectedly, the introduction of the *Phyllobacterium zundukense exsE* gene resulted in intermediate complementation of the Gm resistance phenotype (**Figure 34A**), suggesting that transport of Gm is a broadly conserved function of the SbmA/BacA and related proteins, and is not specific to BclA or SbmA/BacA proteins. As a result, the impact of the nine genes on Gm resistance cannot be used to support the annotation of a protein specifically as BclA or SbmA/BacA; however, it is still a useful metric to test whether a SbmA/BacA-like protein is expressed and functional. Of the six *bclA* genes identified by our screen, three (from *Cyanobacterium aponinum*, *Synechococcus elongatus*, and *Succinivibrio dextrinosolvens*) complemented the Gm resistance phenotype at least as well as the known *bclA* gene of *Bradyrhizobium* sp. ORS285 (**Figure 34A**), confirming they are expressed and functional in *S. meliloti*. The other three *bclA* genes all displayed partial complementation to varying degrees (**Figure 34B**), suggesting they are expressed and functional but either have reduced ability to transport Gm or their expression or stability is sub-optimal. Likewise, the one BclA-like gene (from *Eikenella exigua*) displayed partial complementation of the Gm resistance phenotype (**Figure 34B**). On the other hand, the introduction of the one *bacA* gene that we tested (from *Polymorphum gilvum*) completely failed to complement the Gm resistance phenotype of the *S. meliloti* Δ*bacA*

mutant (**Figure 34B**), which we hypothesize is due to improper expression or stability of the protein rather than functional divergence.

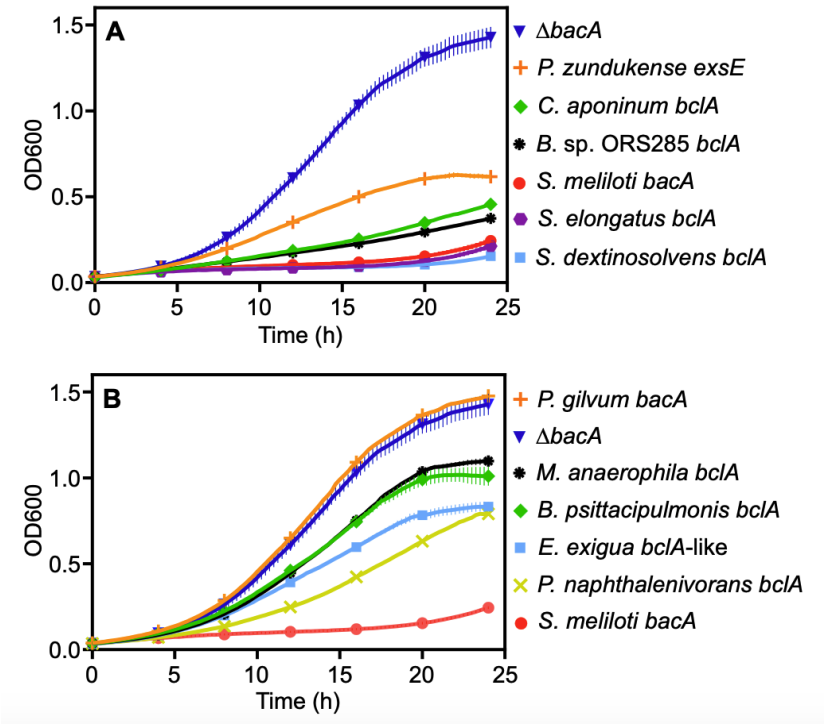


Figure 34 Gentamicin sensitivity assays.

The growth of various *S. meliloti* strains, as measured by OD600, in the presence of 20 µg/mL of gentamicin is shown over a 24-hour period. Each point represents the mean of triplicate wells, with error bars depicting standard deviation. The Δ*bacA* strain represents the *S. meliloti* Δ*bacA* mutant carrying an empty vector, while all other strains are named according to the species of origin of the gene expressed *in trans* in the *S. meliloti* Δ*bacA* background. The experiment was replicated three independent times, and data from a representative experiment is shown. (**A**) Data is shown for genes exhibiting moderate to high level of complementation of the *S. meliloti* Δ*bacA* gentamicin resistance phenotype. (**B**) Data is shown for genes exhibiting low to moderate levels of complementation of the *S. meliloti* Δ*bacA* gentamicin resistance phenotype.

We next indirectly examined whether the nine proteins could transport eukaryotic antimicrobial peptides by measuring the impact of the proteins on the sensitivity of *S. meliloti* to the legume-encoded NCR peptide NCR247 (**Figure 35**); proteins transporting NCR247 are expected to show reduced sensitivity to this peptide. As expected, the *S. meliloti* Δ*bacA* single mutant and the Δ*bacA* Ω*yejA* double mutant were hypersensitive to NCR247 exposure, while the introduction of the known *S. meliloti bacA* or *Bradyrhizobium* sp. ORS285 *bclA* genes *in trans* resulted in reduced sensitivity to NCR247 (**Figure 35**). Introduction of the *P. zundukense exsE* gene into the two mutants resulted in little to no complementation of the NCR247 hypersensitivity phenotypes (**Figure 35**), consistent with the transport of NCR peptides being specific to the SbmA/BacA and BclA family proteins and not a general property of these and

related proteins. All three of the *bclA* genes showing strong complementation of the Gm resistance phenotype (two of which are from cyanobacteria) also showed good complementation of the NCR247 hypersensitivity phenotype (**Figure 35**), confirming the proteins encoded by these three genes are functionally similar to known BclA proteins. In addition, the *bclA* gene from *P. naphthalenivorans* strongly complemented the NCR247 hypersensitivity phenotypes of both strains despite only moderate complementation of the Gm resistance phenotype. Of the remaining two *bclA* genes, one (from *Methylomusa anaerophila*) displayed weak complementation of the NCR247 hypersensitivity (**Figure 35**) and varied in its level of complementation across trials (not shown), while one (from *Basilea psittacipulmonis*) failed to complement (**Figure 35**). Overall, the data for the six BclA proteins support that most BclA proteins are capable of transporting NCR peptides. On the other hand, the NCR247 sensitivity phenotypes of the strains expressing the BclA-like protein from *E. exigua* resembled the phenotypes of the strain expressing *P. zundukense exsE* (**Figure 35**), consistent with BclA-like proteins of Clade B (**Figure 35**) representing a different class of proteins from BclA. In accordance with the Gm resistance data, the *bacA* gene from *P. gilvum* largely failed to complement the NCR247 hypersensitivity phenotypes (**Figure 35**), potentially reflecting improper expression or stability of the encoded protein.

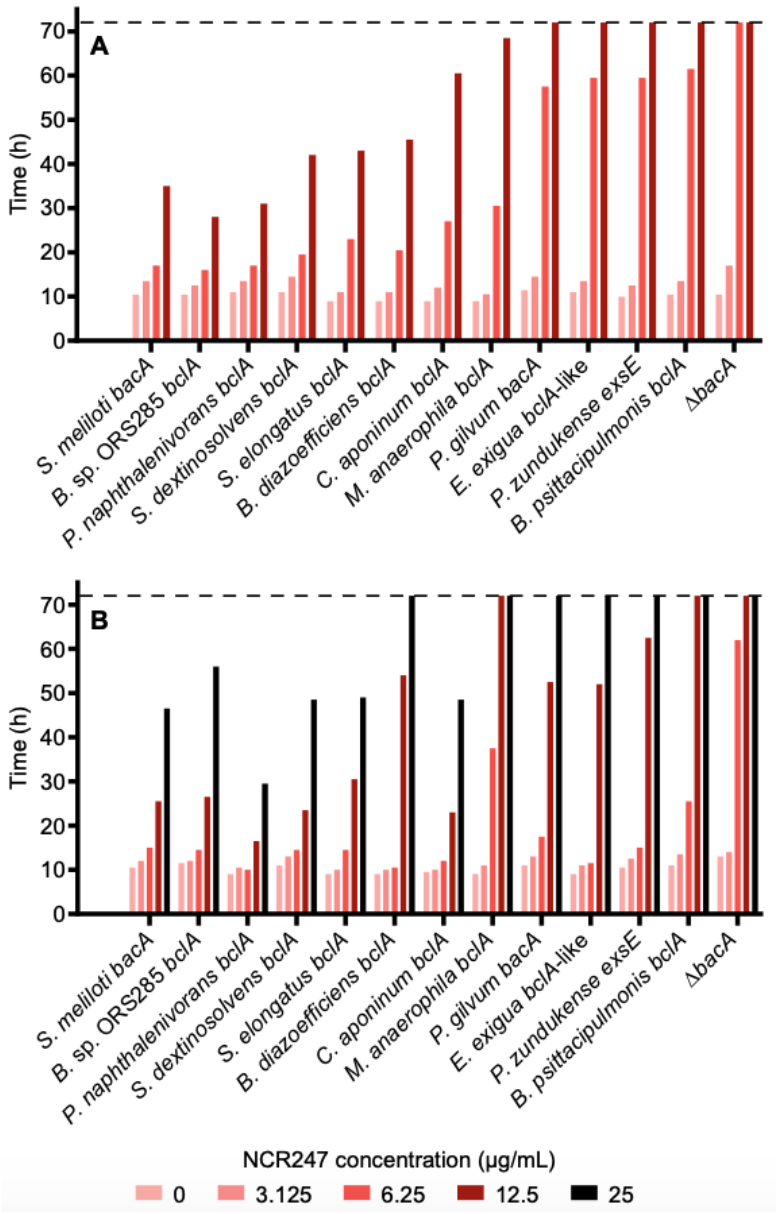


Figure 35 NCR247 sensitivity assays.

The growth of various *S. meliloti* strains, as measured by OD600, in the presence of the antimicrobial peptide NCR247 is shown. Strains were grown in various concentrations of NCR247 as indicated by the shade of red or black. Bars represent the time required for the culture to reach an OD600 of 0.25. Values of 72 hours (indicated by the dashed line) indicate that the strain failed to reach an OD600 of 0.25 within the 72-hour growth period. The Δ*bacA* label represents the *S. meliloti* (A) Δ*bacA* or (B) Δ*bacA* Ω*yejA* mutant carrying an empty vector, while all other strains are named according to the species of origin of the gene expressed *in trans* in the *S. meliloti* (A) Δ*bacA* or (B) Δ*bacA* Ω*yejA* background. (**A**) Data is shown for the *S. meliloti* Δ*bacA* mutant and derivatives. (**B**) Data is shown for the *S. meliloti* Δ*bacA* Ω*yejA* mutant and derivatives.

3. Analysis of the ability of BacA and BclA to support legume symbiosis

We additionally tested whether the nine proteins could complement the nitrogen-fixation defect of a *S. meliloti* Δ*bacA* mutant during symbiosis with *Medicago sativa* (alfalfa) or *Melilotus*

officinalis (yellow-blossom sweet clover). As expected, the *S. meliloti* Δ*bacA* mutant formed small white nodules on both plants and failed to fix nitrogen, while re-introduction of the *S. meliloti bacA* gene *in trans* complemented the nitrogen-fixation phenotype (**Table S1**). All nine of the synthesized genes failed to complement the nitrogen-fixation phenotype (**Table S1**). As the same lack of complementation was observed for the known *bclA* gene of *Bradyrhizobium* sp. ORS285 (**Table S1**), these results suggest that most, if not all, BclA proteins are unable to support an effective symbiosis between *S. meliloti* and its host plants. This is consistent with previous work showing that most *bacA* and *bclA* genes are unable to restore nitrogen fixation when expressed in a *S. meliloti bacA* null mutant (Maruya and Saeki 2010; Guefrachi et al. 2015; Barrière et al. 2017; diCenzo et al. 2017), suggesting that SbmA/BacA and BclA orthologs display slight variations in their peptide substrate range or rate of transport (Huang et al. 2022).

***S. meliloti* genotype ***	**Mean shoot dry weight ± standard deviation (mg/plant)**
***Medicago sativa* (alfalfa)**	
Wildtype Rm1021	66.7 ± 7.1
Δ*bacA* empty vector control	7.6 ± 1.1
Sinorhizobium meliloti 1021 *bacA*	68.6 ± 4.9
Polymorphum gilvum bacA	8.8 ± 1.7
Bradyrhizobium sp. ORS285 *bclA*	7.1 ± 0.6
Succinivibrio dextrinosolvens bclA	7.7 ± 1.1
Basilea psittacipulmonis bclA	10.7 ± 1.2
Polaromonas naphthalenivorans bclA	8.8 ± 0.8
Synechococcus elongatus bclA	6.1 ± 0.9
Methylomusa anaerophila bclA	8 ± 1.4
Cyanobacterium aponinum bclA	7.1 ± 0.2
Eikenella exigua bclA-like	6.7 ± 1.3
Phyllobacterium zundukense exsE	7.7 ± 1.2
Uninoculated control	6.5 ± 1.5
***Melilotus officinalis* (yellow-blossom sweet clover)**	
Wild type Sm1021	72.2 ± 4.5
Δ*bacA*	6.8 ± 1.3
Sinorhizobium meliloti 1021 *bacA*	35.6 ± 8.7
Polymorphum gilvum bacA	5.0 ± 0.6
Bradyrhizobium sp. ORS285 *bclA*	3.1 ± 0.7
Succinivibrio dextrinosolvens bclA	6.3 ± 1.3
Basilea psittacipulmonis bclA	5.2 ± 0.2
Polaromonas naphthalenivorans bclA	3.5 ± 1.9
Synechococcus elongatus bclA	6.8 ± 1.9
Methylomusa anaerophila bclA	6.9 ± 0.6
Cyanobacterium aponinum bclA	4.6 ± 0.2
Eikenella exigua bclA-like	7.2 ± 0.4
Phyllobacterium zundukense exsE	3.6 ± 0.2
Uninoculated	6 ± 1.1

* Strains included the wildtype *S. meliloti* strain Rm1021, an Rm1021 Δ*bacA* derivative, and Δ*bacA* derivatives expressing the genes from the indicated organisms *in trans*.

Table S 2 Shoot dry weights of legumes inoculated with a *Sinorhizobium meliloti* Δ*bacA* mutant and *S. meliloti* Δ*bacA* mutants expressing various sbmA-like proteins in trans.

4. Taxonomic distribution of SbmA/BacA and BclA orthologs across the domain Bacteria

We next examined the taxonomic distribution of the 177 BclA and 71 SbmA/BacA proteins identified as described earlier. Remarkably, 100% and 78% of the identified SbmA/BacA and BclA proteins, respectively, are encoded by species of the phylum *Pseudomonadales* (syn. *Proteobacteria*) (**Figure 36**). As expected, most species encoding SbmA/BacA or BclA proteins encode only one or the other; only six of the 208 species encoding SbmA/BacA and/or BclA encode both, and in all six cases, both genes are carried by the chromosome.

Approximately 70% of the SbmA/BacA proteins are encoded by just two monophyletic groups of organisms, suggesting that SbmA/BacA was acquired at the base of each clade and then vertically transmitted. These two clades are a 24-species clade in the order *Enterobacterales* (all of which encode BacA) and a 29-species clade in the order *Hyphomicrobiales* (25 of which encode SbmA/BacA) (**Figure 36**). Interestingly, the SbmA/BacA proteins of the order *Enterobacterales* form a monophyletic group in the SbmA/BclA protein phylogeny (**Figure S6**). On the other hand, the minimal monophyletic clade encompassing all *Hyphomicrobiales* SbmA/BacA proteins also includes the *Enterobacterales* SbmA/BacA proteins (**Figure S6**). These results suggest that SbmA/BacA proteins of the order *Enterobacterales* were acquired through horizontal transfer from the order *Hyphomicrobiales*. The remaining 22 SbmA/BacA proteins not found within those two clades are distributed across the phylum *Psedomonadales* with no other major clustering observed. Overall, these results suggest that although SbmA/BacA proteins are widespread amongst subclades of the orders *Enterobacterales* (class *Gammaproteobacteria*) and *Hyphomicrobiales* (class *Alphaproteobacteria*), the taxonomic distribution of this protein family is otherwise limited.

BclA proteins show a somewhat broader taxonomic distribution than the SbmA/BacA proteins, although their distribution remains restricted to only a few phyla (**Figure 36**). Like SbmA/BacA, BclA was common in a subclade of the order *Hyphomicrobiales*, in which 19 of 21 species encoded BclA. Most of the other *Alphaproteobacteria* species encoding BclA belong to the order *Rhodospirillales*, in which nine of the 30 species encoded BclA. Within the *Gammaproteobacteria*, the taxon most enriched for BclA proteins was the order *Pasteurellales*, in which 10 of the 16 species encoded BclA. BclA was also abundant in the class *Betaproteobacteria*, unlike SbmA/BacA, and was particularly enriched in the orders *Burkholderiales* (32/60 species) and *Neisseriales* (10/17 species) compared to the orders *Nitrosomonadales* and *Rhodocyclales* (4/27 species across both orders).

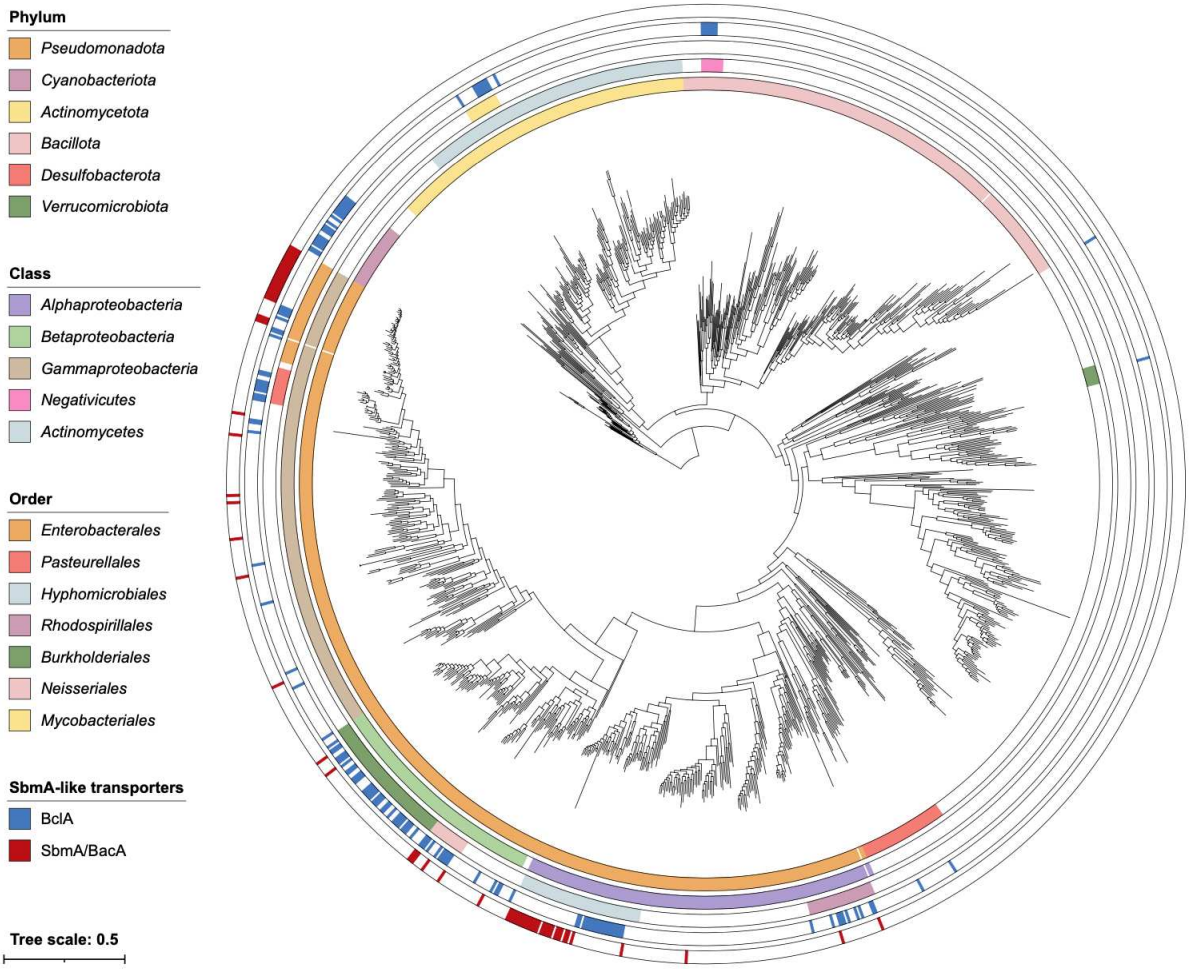


Figure 36 Taxonomic distribution of SbmA/BacA and BclA proteins in the domain Bacteria.

An unrooted maximum likelihood phylogeny of 1,533 bacteria is shown, inferred from the concatenated protein alignments of 31 single-copy proteins. The scale bar represents the average number of amino acid substitutions per site. Three clades of intracellular symbionts/pathogens with long branch lengths were removed for presentation purposes; none of these taxa encode SbmA/BacA or BclA. The outer rings represent the following, starting from the inner ring: (i) the phylum that each strain belongs to, limited to phyla where at least one strain encodes SbmA/BacA or BclA; (ii) the class that each strain belongs to, limited to classes where at least one strain encodes SbmA/BacA or BclA and that are mentioned in the text; (iii) the class that each strain belongs to, limited to classes where at least one strain encodes SbmA/BacA or BclA and that are mentioned in the text; (iv) whether the strain encodes BclA (blue) or not (white); (v) whether the strain encodes SbmA/BacA (red) or not (white). An interactive version of this phylogeny, with node support values and without collapsing of any clades, is provided through iTol (https://itol.embl.de/shared/1IAjjFrHYGLI9), while a Newick-formatted version of the phylogeny can be downloaded from GitHub (https://github.com/amira-boukh/SbmA_BacA_phylogenetic_distribution).

In contrast to SbmA/BacA, which was predicted to be encoded only by species of the phylum *Pseudomonadales*, there were three main clades of organisms predicted to encode BclA outside of the phylum *Pseudomonadales* (**Figure 36**). The largest of these was the phylum

Cyanobacteriota (syn. *Cyanobacteria*), in which BclA was broadly distributed and found in 21 of the 31 species (~67%). The other two main groups of organisms encoding BclA are a subclade of eight species (seven of which encode BclA) of the order *Mycobacteriales* (phylum *Actinomycetota* [syn. *Actinomycetes*]) and the class *Negativicutes* (phylum *Bacillota* [syn. *Firmicutes*]) in which seven of the ten species encode BclA orthologs (**Figure 36**).

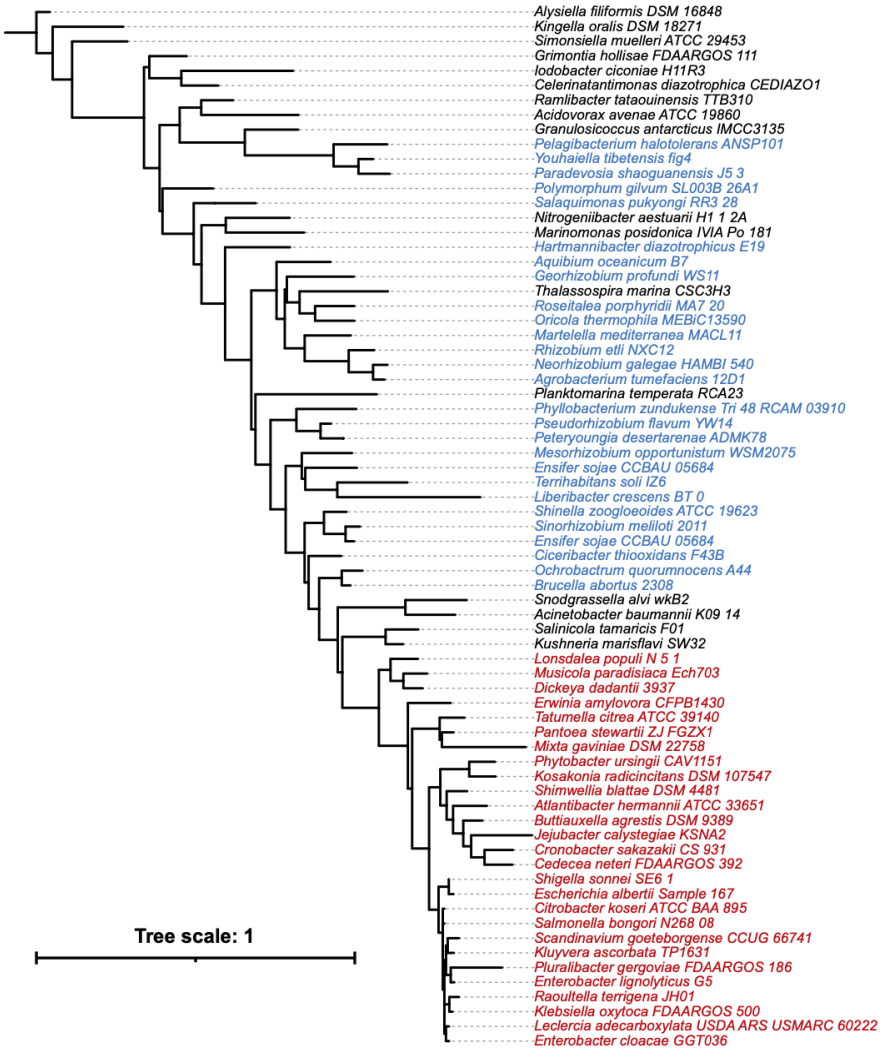


Figure S 6 Maximum-likelihood phylogeny of SbmA/BacA proteins.

A subtree of the maximum likelihood phylogeny of SbmA/BacA-like proteins of Figure 1 is shown. This subtree is limited to the 71 proteins classified as BacA. Proteins encoded by species of the order *Enterobacterales* are shown in red, while proteins of the order *Hyphomicrobiales* are shown in blue. The scale bar represents the average number of amino acid substitutions per site. An interactive version of this phylogeny, with node support values, is provided through iTol (https://itol.embl.de/shared/1IAjjFrHYGLI9).

E. Discussion

We identified 71 SbmA/BacA and 177 BclA orthologs from a search of the proteomes of 1,255 bacterial species. In total, 208 of the 1,255 species (16.6%) encoded at least one copy of SbmA/BacA and/or BclA, with only six of the 208 species (2.9%) encoding both SbmA/BacA and BclA. The observation that SbmA/BacA and BclA proteins were generally not encoded in the same proteome suggests that these protein families have similar biological roles. We also observed that the so-called "*Mycobacterium* BacA" proteins clustered with the BclA proteins in both the SSN and the protein phylogeny, leading us to conclude that the "*Mycobacterium* BacA" proteins are not distinct from BclA; we therefore reclassified the "*Mycobacterium* BacA" proteins as BclA for downstream analyses.

1. Convergent evolution of the SmbA/BacA and BclA protein families

One of the objectives motivating this work was to gain insight into whether the SbmA/BacA and BclA protein families share common ancestry (e.g., that SbmA/BacA evolved from BclA or vice versa) or whether they evolved independently and converged towards a similar function. The taxonomic distribution of SbmA/BacA and BclA proteins within the order *Hyphomicrobiales* is potentially suggestive of the former scenario. Excluding the deep-branching lineages, the order *Hyphomicrobiales* can be sub-divided into two sister clades; SbmA/BacA is widely distributed in one of these clades, while BclA is widely distributed in the other. This could suggest that the SbmA/BacA and BclA proteins of the order *Hyphomicrobiales* evolved from a common ancestral protein present in the ancestor of these clades. However, the *Hyphomicrobiales* SbmA/BacA and BclA proteins are polyphyletic in the BacA/BclA protein phylogeny, which instead suggests that the SbmA/BacA and BclA proteins of the order *Hyphomicrobiales* were independently acquired. The distinct clustering of the BclA and SbmA/BacA proteins in the SSN further supports independent evolutionary origins for these proteins, as does the notably long branch connecting the SbmA/BacA clade to the rest of the phylogeny. Moreover, we consider the differences in transport mechanisms of SbmA/BacA (proton gradient-driven) and BclA (ATP-driven) to be more easily explained if these protein families have separate evolutionary histories. Overall, we interpret the evidence as suggesting that the SbmA/BacA and BclA protein families evolved independently and that their functional similarity is a result of convergent molecular evolution.

Twenty-eight of the BclA proteins were encoded by 21 cyanobacteria. These 28 proteins formed a distinct cluster in the SSN together with 15 non-cyanobacterial BclA proteins, raising the

possibility that these proteins also evolved independently from the rest of the BclA proteins. While we cannot rule out this possibility, we consider the evidence to be insufficient to reach this conclusion at this time.

2. The SbmA/BacA and BclA protein families are associated with eukaryotic host interaction

A second objective of this work was to determine how broadly SbmA/BacA and BclA proteins are distributed across the domain *Bacteria*. Contrary to our initial expectations, we found that both protein families display limited taxonomic distribution. SbmA/BacA orthologs were identified only in the phylum *Pseudomonadales*, with ~89% of the identified BacA proteins being encoded by species of the classes *Alphaproteobacteria* and *Gammaproteobacteria*. A majority of the identified BclA proteins were also found in species of the phylum *Pseudomonadales* with a bias towards the *Betaproteobacteria*; however, BclA proteins were also common in the phylum *Cyanobacteriota*, the class *Negativicutes* (phylum *Bacillota*), and the order *Mycobacteriales* (phylum *Actinomycetota*). Interestingly, many of the clades enriched for species encoding SbmA/BacA or BclA orthologs also include many species known to interact with eukaryotic hosts in mutualistic or pathogenic interactions.

Forty-five of the 55 species (~82%) of the alphaproteobacterial order *Hyphomicrobiales* encode SbmA/BacA and/or BclA; this increases to 45 of 50 species (90%) when excluding the deep-branching *Hyphomicrobiales* lineages. This order accounts for ~79% of the alphaproteobacterial species encoding SbmA/BacA and/or BclA orthologs. Many members of the order *Hyphomicrobiales* are notable for their ability to interact with eukaryotic hosts. All alpha-rhizobia belong to the order *Hyphomicrobiales*, which also encompasses several plant and mammalian pathogens like *Agrobacterium* and *Brucella*, respectively (diCenzo et al., 2023). Similarly, ~75% of the gammaproteobacterial BacA and BclA proteins are encoded by species in the orders *Enterobacterales* and *Pasteurellales*, in which 34 of 47 (~72%; increasing to 81% when excluding a monophyletic group of five obligate endosymbionts) and 10 of 16 (~62.5%) species encode BacA/BclA, respectively. The order *Enterobacterales* is well-known for including many plants (e.g., *Dickeya*, *Pantoea*) and animal/human (e.g., *Klebsiella*, *Yersinia*) pathogens (de la Maza LM, 2020). Likewise, the order *Pasteurellales* encompasses several animal/human pathogens (e.g., *Haemophilus*, *Pasteurella*) (Garrity et al., 2005). In the class *Betaproteobacteria*, BclA and SbmA/BacA were significantly more common in the orders *Burkholderiales* and *Neisseriales* compared to the orders *Nitrosomonadales* and *Rhodocyclales*. The order *Burkholderiales* encompasses all known beta-rhizobia as well as insect gut symbionts

(e.g.., *Caballeronia*) and plant (e.g., *Ralstonia*) and animal/human (e.g., *Burkholderia*) pathogens (Dobritsa & Samadpour, 2016; Voronina et al., 2015). The order *Neisseriales* encompasses many mammalian commensals but also some human pathogens (e.g., *Neisseria*) (Chen et al., 2021).

The phylum *Cyanobacteria* is the largest clade of organisms encoding BclA proteins outside of the phylum *Pseudomonadales*. To our knowledge, cyanobacteria are not pathogenic. However, many can form beneficial associations with diverse hosts, such as the nitrogen-fixing symbiosis between *Nostoc* and plants (Bergman et al., 1992), the mutualistic relationship with fungi (forming lichens), and sponges (Mutalipassi et al., 2021). The order *Mycobacteriales* includes important human and plant pathogens (e.g., *Mycobacterium*, *Rhodococcoides*) (Val-Calvo & Vázquez-Boland, 2023) and opportunistic pathogens (e.g., *Mycolicibacterium*) (Morgado et al., 2022). The class *Negativicutes* is poorly studied despite its peculiar nature, as these *Firmicutes* possess an outer membrane and an LPS (Antunes et al., 2016). Nevertheless, this class is a common component of eukaryotic microbiomes and can cause human disease, including meningitis (Brown, 2016).

The observation that most taxonomic clades enriched for species encoding SbmA/BacA or BclA also contain many mutualistic and/or pathogenic organisms may suggest that eukaryotic host interaction is a driver of SbmA/BacA and BclA maintenance in these lineages. However, the data also suggest that these protein families may pre-date these species interactions. Assuming that SbmA/BacA was acquired by the common ancestor of the SbmA/BacA-containing subclade of the order *Hyphomicrobiales*, the SbmA/BacA protein family potentially evolved in this lineage over 500 million years ago (Rahimlou et al. 2021), which predates the evolution of legumes that are estimated to have evolved around 60 million years ago (Lavin et al. 2005). Thus, SbmA/BacA could not have evolved in this lineage as a response to legume symbiosis. Rather, we hypothesize that SbmA/BacA originally evolved to fulfil another role (such as nutrient transport (Nijland et al. 2024) or protection against membrane-damaging AMPs produced by microbial competitors (Oulas et al. 2021)) and was subsequently co-opted to support legume symbiosis in rhizobia. Likewise, we hypothesize that BclA already existed in the *Bradyrhizobium* lineage prior to the evolution of legume symbiosis, and that this protein was independently co-opted for legume symbiosis in these organisms, mimicking the convergent evolution of NCR peptides in the IRLC and Dalbergioid legume families (Czernic et al. 2015). On the other hand, the absence of SbmA/BacA and BclA proteins in most bacterial lineages may reflect that these proteins also sensitize bacteria to AMPs with intracellular

targets, resulting in a fitness disadvantage in inter- and intraspecific competition. For example, phazolicin is a narrow-spectrum AMP that is produced by some rhizobial strains and that can kill other rhizobia after being imported by BacA and YejABEF transporters (Travin et al. 2023), Likewise, microcins J25 and B17 are narrow-spectrum AMPs produced by *E. coli* that depend on SbmA/BacA to reach their intracellular targets in target organisms (Parker and Davies 2022). Considering this, we hypothesize that SbmA/BacA and BclA have been selected for in bacteria where resistance to membrane-targeting AMPs is more important than resistance to AMPs with intracellular targets, such as during host interaction, where these proteins may have been repeatedly co-opted to help bacteria survive exposure to host-encoded AMPs. In contrast, we hypothesize that the absence of these proteins is favoured when bacteria predominately encounter AMPs with intracellular targets, which may be the case for non-host associated microbes primarily encountering AMPs produced by other microbes.

3. Transport of AMPs is a general property of SbmA-like protein

The abilities of several newly identified BclA proteins to complement the phenotypes of an *S. meliloti* Δ*bacA* mutant were tested to validate that these proteins were correctly annotated. *S. meliloti bacA* null mutants display increased gentamicin resistance compared to the wild type (Ichige & Walker, 1997). Eight of the nine synthesized genes at least partially complemented the gentamicin resistance phenotype of a *S. meliloti* Δ*bacA* mutant, suggesting these eight proteins were expressed and at least partially functional in *S. meliloti*. Interestingly, even the gene encoding an ExsE ortholog partially complemented the gentamicin resistance phenotype, indicating that gentamicin transport is not specific to SbmA/BacA and BclA proteins but is a general property of these and related protein families. Gentamicin sensitivity assays are commonly used to characterize the function of rhizobial *bacA* orthologs and rhizobial *bacA* mutant alleles generated through site-directed mutagenesis (LeVier & Walker, 2001). Although these assays are useful to identify null phenotypes, our results show that they do not probe a function unique to SbmA/BacA or BclA proteins and thus have limited value as a proxy to peptide transport or host interaction assays.

In addition to showing increased resistance to gentamicin, *S. meliloti* Δ*bacA* mutants show increased sensitivity to NCR peptides (Haag et al., 2011; LeVier & Walker, 2001). As the antimicrobial activity of NCR peptides is a result of their interaction with the cell envelope, it is thought that SbmA/BacA and BclA proteins provide resistance to NCR peptides by moving the peptides away from the cell envelope and into the cell (diCenzo et al., 2017; Haag et al.,

2011). SbmA/BacA and BclA proteins have also been shown to transport other antimicrobial peptides, including mammalian antimicrobial peptides such as Bac7 (Domenech et al., 2009; Guefrachi et al., 2015; Marlow et al., 2009; Runti et al., 2013). As expected, only the proteins annotated as BclA were capable of effectively complementing the sensitivity of *S. meliloti* Δ*bacA* and *S. meliloti* Δ*bacA* Ω*yejA* mutants to the NCR peptide NCR247. Of the six newly identified BclA proteins that were tested, four repeatedly demonstrated good levels of complementation; these proteins were from *P. naphthalenivorans* (class *Betaproteobacteria*), *S. dextrinosolvens* (class *Gammaproteobacteria*), *S. elongatus* (phylum *Cyanobacteriota*), and *C. aponinum* (phylum *Cyanobacteriota*). The other two, from *M. anaerophila* (class *Negativicutes*) and *B. psittacipulmonis* (class *Betaproteobacteria*), showed weak and variable or little to no complementation, respectively. However, there are thousands of distinct NCR peptides encoded across the legume family (Montiel et al., 2017), and thus, the inability of a transporter to transport NCR247 does not mean that it is unable to transport other NCR peptides or mammalian antimicrobial peptides. Indeed, *S. meliloti yejA* mutants show increased sensitivity to the peptide NCR280 but not NCR247 (Nicoud et al., 2021). Regardless, these results support that the ability to transport antimicrobial peptides, including NCR peptides, is a general property of bacterial SbmA-like proteins.

F. Conclusion

In summary, we identified 208 bacterial species encoding SbmA/BacA or BclA. These species were not equally distributed across the domain *Bacteria*; instead, SbmA/BacA proteins were found only in the phylum *Pseudomonadota*, while BclA proteins were primarily found within a subset of families across four phyla. Our analyses suggest that the SbmA/BacA and BclA protein families arose independently and that their functional similarity is a result of convergent evolution rather than shared ancestry. Our data also support the hypothesis that SbmA/BacA and BclA proteins have been repeatedly co-opted to facilitate both mutualistic and pathogenic associations with eukaryotic hosts by allowing bacteria to cope with host-encoded antimicrobial peptides. We further suggest that the distribution of SbmA/BacA and BclA is determined by the fitness trade-off of their presence. Specifically, we predict that genes encoding SbmA/BacA or BclA will only be maintained in bacteria for which resistance to membrane-targeting AMPs is more important than resistance to AMPs with intracellular targets.

G. Materials and methods

1. Bacterial strains and growth conditions

The bacterial strains used in this study are listed in **Table S2**. *E. coli* strains were cultured at 37 °C using Lysogeny Broth (LB; 10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl). *S. meliloti* strains were grown at 28 °C using either LBmc (LB supplemented with 2.5 mM $CaCl_2$ and 2.5 mM $MgSO_4$), YEB (0.5% beef extract, 0.1% yeast extract, 0.5% peptone, 0.5% sucrose, 0.04% $MgSO_4$ $7H_2O$, pH 7.5), or MM9 minimal medium (2% MOPS-KOH, 1.92% NH_4Cl, 0.35% NaCl, 0.2% KH_2PO_4, 0.2% $MgSO_4$, 0.05% $CaCl_2$, 0.05% Biotin, 0.0004% $CoCl_2$, 0.38% $FeCl_3$, 1% Glucose, 1% Na_2-succinate). Antibiotics were added as appropriate and included: ampicillin (Amp; 100 µg/mL), kanamycin (Km; 100 µg/mL), streptomycin (Sm; 200 or 500 µg/mL), spectinomycin (Sp; 50 µg/mL), and tetracycline (Tc; 5 µg/mL). Antibiotic concentrations were generally halved for liquid cultures.

2. Cloning of bacA, bclA, and exsE homologs

Ten vectors encoding putative *bacA*, *bclA*, or *exsE* genes, codon optimized for *S. meliloti* 1021 and flanked by XbaI and BamHI recognition sites, were produced by Twist Biosciences (**Table S2, Dataset S1**). Each gene was PCR amplified from the plasmids using Q5 polymerase (New England Biolabs; NEB) with the primers 5'-GAAGTGCCATTCCGCCTGACC and 5'-CACTGAGCCTCCACCTAGCC. The resulting amplicons were individually digested with XbaI/BamHI and ligated into XbaI/BamHI-digested expression vector pRF771 (Wells & Long, 2002). Plasmids were sequence verified via Illumina sequencing (151 bp paired-end reads) at SeqCenter (Pittsburg, PA, USA), after which reads were aligned to the expected template sequences using bowtie2 version 2.5.0 (Langmead & Salzberg, 2012) and alignments visualized using the Integrative Genomics Viewer version 2.12.3 (J. T. Robinson et al., 2011).

3. Transfer of plasmids to *S. meliloti*

All plasmids of interest were transferred to a *S. meliloti* Δ*bacA* mutant via triparental mating using the helper strains *E. coli* MT616 or *E. coli* HB101, as described previously (Barrière et al., 2017; Finan et al., 1986). Transconjugants were recovered through plating of mating spots on LBmc Sm^{200} Tc or YEB Sm^{500} Tc plates. Likewise, plasmids were transferred to a *S. meliloti* Δ*bacA* Ω*yejA* double mutant via triparental mating as described previously (Barrière et al., 2017), with transconjugants recovered on YEB Sm^{500} Tc Km Sp plates. All transconjugants were streak purified three times prior to use.

4. Gentamicin sensitivity assays

Gentamicin sensitivity assays were performed largely as described previously (diCenzo et al., 2017). Briefly, overnight cultures of *S. meliloti*, grown in LBmc Sm^{100} Tc, were washed and resuspended in LBmc to an optical density at 600 nm (OD600) of 1.0. Ten µL aliquots of the cell suspensions were added to triplicate wells of a 96-well plate and mixed with 190 µL of LBmc with or without 20 µg/mL of gentamicin (Gm). A Gm concentration of 20 µg/mL was chosen for the assays based on preliminary sensitivity assays (**Figure S7**). Plates were tape-closed to prevent evaporation and then incubated at 30˚C with maximal shaking in a BioTek Synergy H1 plate reader for 24 hours. OD600 measurements were collected every 15 minutes using the Gen5 software (Agilent Technologies).

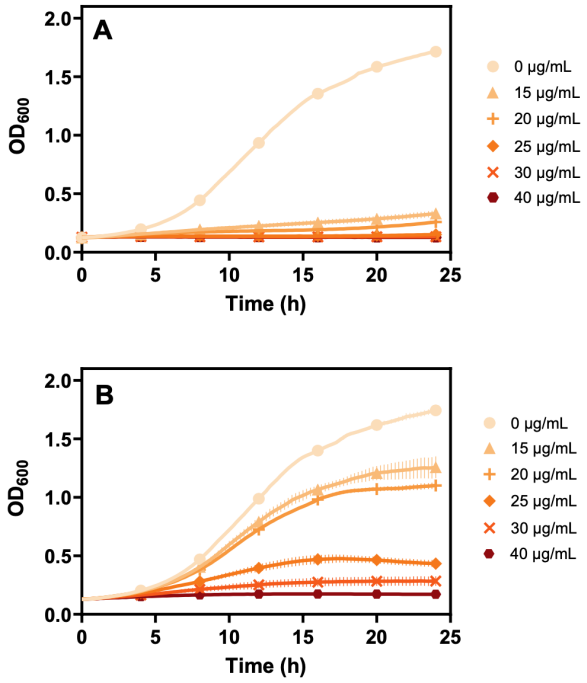


Figure S 7 Effect of gentamicin on the growth of Sinorhizobium meliloti

Growth of *S. meliloti* in LBmc containing various concentrations of gentamicin, as indicated in the legend. Growth profiles are shown for (**A**) *S. meliloti* Δ*bacA* carrying an empty expression vector and (**B**) *S. meliloti* Δ*bacA* expressing the *S. meliloti bacA* gene *in trans*. Each point represents the mean of triplicate wells, with error bars depicting standard deviation. The experiment was replicated three independent times, and data from a representative experiment is shown.

5. NCR247 sensitivity assays

NCR sensitivity assays were performed largely as described previously (Nicoud et al., 2021). Briefly, overnight cultures of *S. meliloti*, grown in MM9 minimal media, were washed and resuspended in MM9 to an OD600 of 1.0. Cell suspensions were then diluted to an OD600 of

0.05 and 145 µL transferred to the wells of 96-well plates and mixed with 5 µL of an NCR247 solution to reach final concentrations of 50, 25, 12.5, 6.25, 3.125, and 0 µg/mL of NCR247. Plates were incubated at 28˚C with shaking (180 rpm) in a Tecan Spark plate reader for 72 hours, and OD_{600} measurements were taken every 30 minutes and processed using the SparkControl software (Tecan).

6. Plant symbiotic assays

Seeds of *M. sativa* cv. Algonquin (alfalfa) and *M. officinalis* (yellow-blossom sweet clover) (Speare Seeds Limited; Harriston, Ontario, Canada) were surface-sterilized and germinated on water agar plates for two nights in the dark, as described previously (Huang et al., 2022). Leonard assemblies were prepared as described before (13), with a 1:1 (w/w) mixture of vermiculite and silica sand in the top pot, 250 mL Jensen's medium (Jensen & Jensen, 1942) in the bottom pot, and a cotton wick connecting the pots, and then autoclaved. Five seedlings were sown per pot, and assemblies were incubated for two nights. Assemblies were next inoculated in triplicate with 1 x 10^8 CFU of *S. meliloti* per assembly. Plants were grown in a Conviron growth chamber with an 18-hour photoperiod, 300 µmol/s of light, 21 ºC daytime temperature, and 17 ºC nighttime temperature. After 30 days, plant shoots were collected and dried at 60 ˚C for six nights prior to weighing.

7. Phylogenetic analysis of BacA and BclA proteins

GenBank files corresponding to 3498 RefSeq bacterial genomes with 'complete' genome assemblies were downloaded from the National Center for Biotechnology Information (NCBI) Genome Database. A subset of the genomes was prepared by collecting genomes from one representative genome per genus, using the genome from the first species per genus when sorted alphabetically. The phylogenetic analyses were then repeated twice: once using all 3498 RefSeq bacterial genomes and once using the reduced set of 1255 genomes (**Dataset S2**). As the results were similar, we only present results generated using the reduced dataset.

BacA, BclA, and related proteins were extracted from the bacterial proteomes using a modified version of an existing in-house pipeline (diCenzo et al., 2019). The seed alignment of the SbmA/BacA-like family, consisting of eight sequences, was downloaded from PFAM (PF05992), and a hidden Markov model (HMM) built using the hmmbuild function of HMMER version 3.3 (Johnson et al., 2010). Separately, a HMM database was built by combining (i) the complete PFAM-A version 31.0 HMM database, (ii) the complete TIGERFAM version 15.0 HMM database, (iii) HMMs built from the seed alignments of PRK11098 (105 sequences in

the seed alignment) and COG1133 (nine sequences in the seed alignment) downloaded from NCBI's Conserved Domain Database, and (iv) HMMs built for each of the BacA (15 sequences), BclA (5 sequences), *Mycobacterium* BacA (10 sequences), ExsE (6 sequences), and *Bradyrhizobium* homologous clade (7 sequences) proteins used in the phylogenetic analysis of (Guefrachi et al., 2015). Next, the hmmsearch function of HMMER was used to search all bacterial proteomes using the PF05992 (SbmA/BacA-like family) HMM. All hmmsearch hits were then scanned against the full HMM database using the hmmscan function of HMMER. Each protein was annotated according to the top-scoring HMM from this search.

Proteins annotated as BacA, BclA, *Mycobacterium* BacA, ExsE, or *Bradyrhizobium* homologous clade were extracted and aligned using Clustal Omega version 1.2.4 (Sievers et al., 2011), hmmalign from HMMER (Johnson et al., 2010) and MAFFT version 7.45 (Katoh & Standley, 2013) and alignment quality assessed with T-COFFEE version 13.45 (Notredame et al., 2000). Poor quality regions of the best scoring alignment (Clustal Omega) were removed using trimAl version 1.4 with the automated1 option (Capella-Gutiérrez et al., 2009) and then used as input for maximum likelihood phylogeny inference using IQ-TREE2 version 2.2.0 (Minh et al., 2020) with the LG+F+I+I+R9 model. The LG+F+I+I+R9 model was used as it was identified as the best-scoring model by the IQ-TREE2 implementation of ModelFinder (Kalyaanamoorthy et al., 2017) based on Bayesian information criterion (BIC), with model search limited to the LG, WAG, JTT, Q.pfam, JTTDCMut, DCMut, VT, PMB, BLOSUM62, and Dayhoff models. Branch supports were assessed in IQ-TREE using a Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010) and an ultrafast bootstrap analysis, with both metrics calculated from 1000 replicates. All phylogenies created in this study were visualized with the iTOL web server (Letunic & Bork, 2021).

8. Sequence similarity network analysis

A sequence similarity network (SSN) was constructed for the 366 proteins identified using the HMM approach described above. The SSN was constructed using the online Enzyme Function Initiative's Enzyme Similarity Tool (EFI-EST; https://efi.igb.illinois.edu/efi-est/) (Oberg et al., 2023; Zallot et al., 2019) with an alignment score threshold of 115, corresponding to an approximate sequence ID ≥ 35%. The resulting network was visualized using Cytoscape version 3.10.1 (Shannon et al., 2003).

9. Multilocus sequence analysis

A bacterial species phylogeny was produced for the 1,253 representative bacterial species using

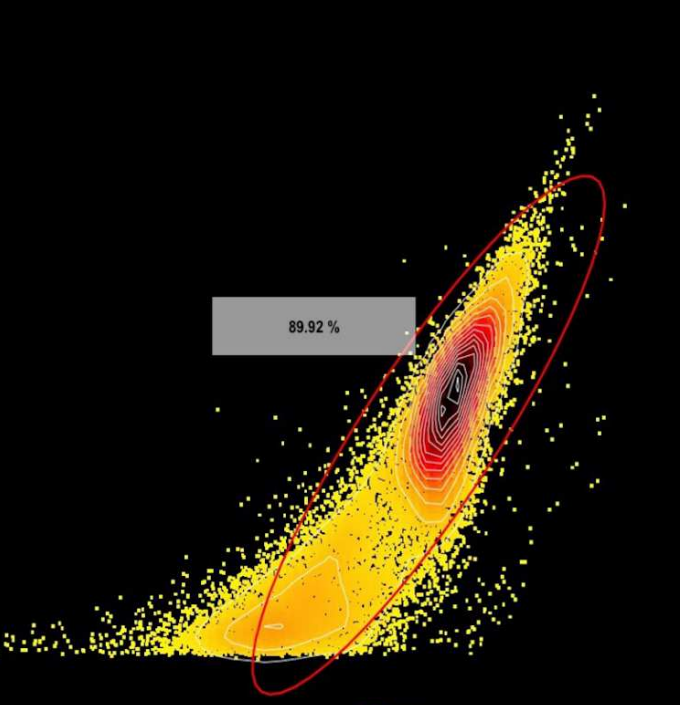
an adaptation of an existing in-house pipeline (diCenzo et al., 2019); two of the 1,255 downloaded genomes were excluded as they encoded none of the marker genes. First, orthologs of 31 highly-conserved, single-copy proteins (DnaG, Frr, InfC, NusA, Pgk, PyrG, RplA, RplB, RplC, RplD, RplE, RplF, RplK, RplL, RplM, RplN, RplP, RplS, RplT, RpmA, RpoB, RpsB, RpsC, RpsE, RpsI, RpsJ, RpsK, RpsM, RpsS, SmpB, Tsf) were identified in the 1,253 proteomes using the AMPHORA2 pipeline (Wu & Scott, 2012). Each group of orthologs was individually aligned using MAFFT (Katoh & Standley, 2013) and trimmed using trimAl (Capella-Gutiérrez et al., 2009). The protein alignments were then concatenated and used as input for ModelFinder as implemented in IQ-TREE2, and the best scoring model was identified based on BIC. IQ-TREE2 was then used to infer a maximum likelihood phylogeny from the concatenated alignment using the LG+I+I+R10 model. Branch supports were assessed in IQ-TREE using the Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT) [22] and ultrafast jackknife analysis with a subsampling proportion of 40%, with both metrics calculated from 1000 replicates.

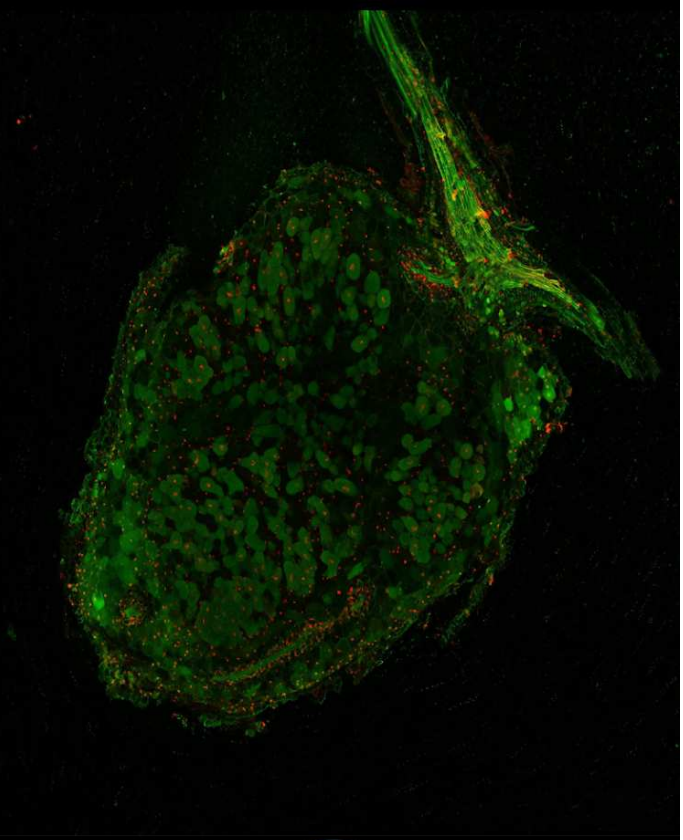
H. Data availability

All genome sequences used in this work were previously published, and the assembly accessions are provided in Dataset S2 (N. T. Smith et al., 2024). Likewise, all protein sequences included in Figure 33 are provided in Dataset S1 (N. T. Smith et al., 2024). Newick formatted phylogenies used to create Figures 33 and 36 are available through GitHub (https://github.com/amira-boukh/SbmA_BacA_phylogenetic_distribution). All code to repeat the analyses in this study is also available through GitHub (https://github.com/amira-boukh/SbmA_BacA_phylogenetic_distribution).

I. Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants to G.C.D. and G.W.H. N.T.S. was supported, in part, by an R. S. McLaughlin Fellowship from Queen's University and a Wicked Ideas grant from Queen's University to G.W.H and G.C.D.. A.B. benefited from a Ph.D. contract in the frame of the CNRS 80|PRIME – 2021 program and was partially supported by a Mitacs Globalink Research Award. B.A. benefited from a French State grant (Saclay Plant Sciences, reference n° ANR-17-EUR-0007, EUR SPS-GSR) under a France 2030 program (reference n° ANR-11-IDEX-0003).





Discussion

A. The distribution of NCR peptides across legume species and clades

Legume species from five legume clades have evolved independently multiple times the ability to induce Terminal Bacteroid Differentiation (TBD) in their rhizobial host, suggesting a significant adaptive advantage for the host plants (Oono et al., 2010). However, NCR-induced TBD has been reported only in two of those clades, IRLC and Dalbergioids, and their evolution remained undeciphered (Czernic et al., 2015; Montiel et al., 2017). NCR peptides were first discovered in IRLC species (*Medicago truncatula*) (Mergaert et al., 2003), and it was assumed for a long time that NCR peptides were specific to this clade. However, the discovery of NCR peptides in the Dalbergioid species (Czernic et al., 2015; Raul et al., 2021) showed that this was not true and suggested that in the five clades, TBD is induced by NCR peptides. Still, the presence of these peptides in the other TBD-inducing clades was unexplored. Moreover, different evolutionary theories were suggested concerning the evolution of NCR peptides. Some studies suggested that NCR peptides have evolved independently in IRLC and Dalbergioids (Czernic et al., 2015), supporting the convergent evolution of TBD (Oono et al., 2010). However, another study demonstrated by sequence-based phylogenetic analysis that NCR peptides and defensins may share the same evolutionary origin (Salgado et al., 2022). However, this study used only a small subset of type-2 NCR peptides.

As presented in the results part, here, we showed that *Indigofera argentea* (Indigoferoids clades) induce TBD of *Bradyrhizobium elkanii* bacteroids that showed a polyploid genome. Moreover, we found NCR peptides in the three studied *Lupinus* species, suggesting that the TBD observed in the Genistoids clade is also induced by NCR peptides. Furthermore, an old study about the *Lupinus-Bradyrhizobium* symbiosis (Fernández-Pascual et al., 2007), based on nodules light microscopy images, suggested that no TBD occurred in *Lupinus albus* (non-swollen bacteroids), while another study of *Lupinus angustifolius* (Dart & Mercer, 1966*)* demonstrated that TBD occurs in this species. However, with our approach, we found highly expressed NCR peptides in both *Lupinus albus* and *Lupinus angustifolius.* Nevertheless, the number and the expression of NCR peptides were higher in *Lupinus angustifolius* than in *Lupinus albus.* Taking into consideration these findings, we suggest that both Lupinus species induce TBD, but with different degrees, where *Lupinus angustifolius* has more differentiated (swollen) bacteroids. Indeed, it has been demonstrated previously that there is a positive correlation between the degree of TBD and the number of NCR peptides (Montiel et al., 2017). Consistently, *I. argentea,* which induced a relatively lower TBD level with bacteroids displaying 3C of DNA compared to 24C in *M. sativa*, contained only 12 NCR peptides.

NCR peptides were present in the four studied TBD-inducing, supporting the suggestion that TBD is induced by NCR peptides. It has been shown previously with blast analysis that in legume plants that do not induce TBD, such as *Lotus japonicus* and *Glycine max*, NCR genes are absent. However, according to our computational analysis, few putative NCR genes are present in *Lotus japonicus, Glycine max, Phaseolus vulgaris,* and *Vigna angularis.* However, they were not expressed or barely expressed in the nodules of these species that do not induce TBD, which validates the involvement of NCR peptides in the TBD process and suggests a common origin of NCR peptides. However, these putative peptides had different 3D structures than NCR peptides. Still, two of these 3D structures cluster with NCR peptides in the superclusters. Thus, whether these peptides are really NCR peptides and they do not induce TBD because of the lack of other essential NCRs or they are not NCR peptides remains to be elucidated. Different hypotheses could be suggested about the absence of NCR peptides in some legume clades. We know that legume plants induce TBD because it increases the symbiotic efficiency by increasing nitrogen fixation. Thus, it is possible that the legumes that do not induce TBD did not live in nitrogen starvation conditions in nature, and so, it was not advantageous to evolve this trait. However, it could also be a random evolutionary effect. Furthermore, the numbers, the sequences, and the structures of NCR peptides were highly variable among legume species that induce TBD and positively correlated with the amount of TBD. This suggests that in some species, diversification and expansion of the NCR family occurred, while in other legume species, this family did not expand. The more probable hypothesis of these differences is related to the BacA transporter of their symbionts being capable of transporting NCR peptides with specific characteristics. Consistently, it has been shown that *S. fredii* and *S. leguminosarum* can support symbiosis with *P. sativum* and *M. officinalis* but not with *M. sativa,* which have more NCR peptides and more cationic ones (diCenzo et al., 2017). Moreover, it has been demonstrated that BclA from *Bradyrhizobium sp.* ORS285 can partially complement the function of *S. meliloti* BacA and thus supports NCR peptides of *M. sativa*, while the inverse was not true, suggesting that *S. meliloti* BacA may not be capable of transporting all *Aeschynomene* plants NCRs (Guefrachi et al., 2015).

Studying the orthology and clustering of the known NCR peptides in IRLC and Dalbergioids validated that NCR peptide sequences are highly different in these two clades and that they may have taken different evolutionary pathways to induce TBD. However, extending the study to two other clades and other new NCR peptides in the four clades demonstrated that NCR peptides were not clade-specific where different clusters regroup IRLC and Genistoid NCRs,

while Dalbergioid NCRs regroup with Indigoferoid's ones. The presence of type-1 and type-2 NCRs in Indigoferoids and Dalbergioids validated the rapid divergence of NCR peptides in each species and suggested that NCR peptides from these clades may have evolved from defensins. Yet, their sequences were not homologous to defensins.

Furthermore, it has been recently reported that NCR peptides in *Medicago truncatula* did not have specific transcription factors, but they were induced by AHL transcription factors (S. Zhang et al., 2023). AHLs are DNA-binding proteins belonging to the Type I AT-Hook Motif Nuclear Localized (AHL) transcription factor family (S. Zhang et al., 2023). It has been shown that AHL transcription factors are conserved in non-IRLC species and can induce the expression of the essential peptide in *M. truncatula* NCR169 (S. Zhang et al., 2023). Thus, we wonder whether the same transcription factors induce NCR peptides in other species and clades that induce TBD or not.

When we clustered the known NCR peptides from IRLC and Dalbergioids, a lot of NCR peptides were regrouped into clusters with other proteins. As expected, those proteins were no-annotated NCR peptides, of which most of them were recovered with SPADA analysis. Moreover, when we searched NCR peptides with the SPADA pipeline, after filtering out the non-NCR peptides, we searched what gene families belong to those peptides. Blast analysis of these filtered-out peptides against the NCBI NR (Non-Redundant) database highlighted the presence of LTP (Lipid Transfer Protein) and defensin proteins. However, only defensin sequences were used for the structure-based analysis that showed that NCR peptides from Dalbergioids and Indigoferoids evolved parallelly from defensins, while a convergent evolution has driven the evolution of NCR peptides in IRLC and Dalbergioids with no evidence about the ancestral gene family. Therefore, even though 3D structures of LTPs are different from NCR peptides with only alpha helices, while NCR peptides seem to have one alpha-helix and 2 beta sheets with some variations, supplementing our dataset with LTP sequences and test if they might participate in the evolution of NCR peptides would be interesting.

The functional experiments further allowed us to validate our approach of searching NCR peptides where seven from the nine selected NCR peptides induced TBD features in *S. meliloti* free-living bacteria. Among these peptides are new NCRs in the well-studied model legumes such as *M. truncatula* and new NCRs in the unstudied clades such us Indigoferoids. These results highlights the robustness of our approach to find NCR peptides.

Furthermore, the SPADA pipeline allowed us to efficiently predict new NCR peptides in all studied legume species, which highly outperforms the blast searches. SPADA is a specialized tool for predicting small cysteine-rich peptides in plant genomes. It has demonstrated high sensitivity and specificity in identifying peptide families' rich in cysteine (P. Zhou et al., 2013). Specifically, SPADA is distributed with a prediction model tailored for *M. truncatula*, further improving its efficiency in predicting NCR peptides (P. Zhou et al., 2013). However, the prediction capabilities of SPADA may have limitations when it comes to specific families of cysteine-rich peptides that can have unique structural motifs and disulfide bonding patterns that may not align perfectly with the general characteristics of cysteine-rich peptides typically predicted by SPADA. Indeed, since SPADA is tailored for *M. truncatula* and 90% of our NCR clusters were from IRLC, our analysis predicted more NCR peptides in the IRLC and Genistoids clades that were homologous. Even though the amount of TBD was consistent with the number of NCR peptides found, it is possible that we missed some Indigoferoids and Dalbergioids NCRs because of the sequence and structural differences of their NCRs compared to what SPADA was tailored for. In summary, while SPADA is a powerful tool for predicting cysteine-rich peptides in general, we may need to complement SPADA with additional tools or customized approaches tailored to the distinct characteristics of these peptide families. For example, the use of machine learning approaches to predict NCR peptides could be a complementary analysis to the sequence-based statistical approaches (SPADA) we used to search for NCR peptides (Klimovich & Bosch, 2024).

B. Deciphering the evolution of NCR peptides using structural phylogenetics

Amino acid sequences of NCR peptides could not provide sufficient information about the evolution of these peptides. Although tailored sequence-based approaches, such as SPADA (P. Zhou et al., 2013), allowed us to find new NCRs and classify them, they did not clear up the evolutionary and functional information. The computational structural approaches used provided more insights about the NCR peptides through the prediction of the 3D structures of the known NCR peptides, the new ones, and the defensins family. The comparison and clustering of these 3D structures further extend the evolutionary context to a broader scale and highlight important evolutionary information that the clade-specific sequence clusters could not uncover.

First, our study highlights a complex evolutionary history of NCR peptides with the presence of convergent, parallel, and divergent evolution. On one hand, we suggested that NCR peptides

from Dalbergioids and Indigoferoids evolved parallelly from defensins (**Figure 37**). On the other hand, a convergent evolution may drive the evolution of IRLC and Genistoids NCR peptides that were similar, whereas their clades are relatively distant in the legume phylogeny, and no evidence about their ancestral gene family was found (**Figure 37**). Taking into consideration these results and the few non-expressed putative NCR peptides in the legumes that do not induce TBD having highly diverse 3D structures from NCR peptides, the above suggestion of a common origin of NCR peptides was rejected. At the inter-clade scale, we believe that a defensin protein shared in the common ancestor of legume species evolved at least two times to form sequence-unrelated structurally similar NCR peptide families in Dalbergioids and Indigoferoids. (**Figure 37**) Whether NCR peptides from IRLC and/or Dalbergioids also evolved from defensins and lost their structure similarity remains unknown, where these peptides had similar sequences and structures between each other but were different from all other NCRs and defensins. A large scale study with more legume species and clades may be required to resolve the ancestral origin of these peptides. At the species scale, another event that plays an important role in the evolution of NCR peptides is the recent duplication events followed by rapid diversification where, in each legume species, different NCR peptides that are in the same genomic region regroup in the same orthologous clusters. Indeed, it has been suggested previously that the expansion of NCR peptides is driven by gene duplication followed by diversification (Alunni et al., 2007; Montiel et al., 2017; Zorin et al., 2022). After a duplication event of one ancestral NCR peptide, a diversification occurs between the paralogs, and they lose their sequence identity while maintaining or not a detectable structural similarity, giving rise to different superclusters and species-specific NCR peptides.

We suggest that the sequence-unrelated structurally similar NCR peptides in one legume species may have different functions during the symbiosis. Indeed, it has been previously shown that NCR peptides act in waves during different stages of nodule formation and bacteroid differentiation (Guefrachi et al., 2014). For instance, it has been shown that the NCR343 in *M. truncatula* is active in the infected cells of IZ and ZIII nodule zones, while the NCR-new35 could be detected only in IZ cells (Horváth et al., 2023). Therefore, the two tested NCR peptides from *A. sinicus* and *T. pratense* that did not induce polyploidy in *S. meliloti* do not show that these peptides are not essential. They may act in combination with other NCR peptides or have another function than interacting with cell cycle regulators to induce polyploidy. Indeed, it has been previously suggested that some NCR peptides are involved in the maintenance of the bacteroid state rather than inducing TBD and thus formation the bacteroid (Mergaert, 2018). In

this context, the emergence of sequence-unrelated NCR peptides could be the consequence of functional divergence. We suggest that the disulfide bounds restrict the NCR diversification in the same clade into a structurally limited landscape, which allow the diversification of NCR peptides while maintaining a detectable structure similarity.

Through this study, we demonstrated the importance of combining sequence and structure-based computational approaches with functional experiments to decipher complex evolutionary histories.

C. BacA and BclA proteins are independently co-opted for NCR resistance in rhizobia

In order to cope with NCR peptides, the rhizobial symbionts uses their ABC transporters called BacA or BclA. Our study allowed us to decipher the evolutionary history of BacA/BclA transporters at the bacterial domain scale. The first result was that 202 of the 208 bacterial species that encodes BacA/BclA genes, encoded only one of them validating that these two proteins have the same function. Our results based on phylogenetic analysis, sequence similarity network and functional assays allowed us to suggest that the functional similarity of these transporters arise from convergent evolution (**Figure 37**).

Some phylogenetic features suggest that BacA and BclA share a common ancestor. For instance, the BacA clade belongs to the BclA clade, and a common ancestor is shared between two clades of the order of *Hyphomicrobiales* in the species tree, of which BacA is widely distributed in one clade while BclA is distributed in the other. However, these BacA and BclA from *Hyphomicrobiales* order are polyphyletic in the BacA/BclA tree, the BacA and BclA share different sequence clusters, a long branch links the BacA clade to the rest of tree, BacA and BclA have different 3D structures and different mechanism of transport. All together, these results allowed us to suggest that BacA and BclA evolved independently (**Figure 37**).

The analysis of the distribution of BacA and BclA at a broader scale across the bacterial domain allowed us to find BacA and BclA only in a limited taxon, of which BclA had more broad distribution. Moreover, these taxonomic clades enriched for BacA/BclA contained many mutualistic and/or pathogenic organisms, which suggested that the host-microbe interactions may drive the maintenance of these transporters in these lineages. However, the evolution of Bacteria predates their interactions with these species, where BacA probably evolved 500 million years ago in the *Hyphomicrobiales* while BclA already existed in the Bradyrhizobium

lineage, both prior to the evolution of legume 60 million years ago, suggesting that BacA and BclA may evolved from another function and independently co-opted from legume-rhizobia symbiosis to cope with NCR peptides, mimicking the convergent evolution of NCR peptides (**Figure 37**).

The functional experiments allowed us to validate our identified putative BacA/BclA transporters and also our phylogenetic assumptions. Indeed, all the tested proteins were capable of transporting gentamicin which validates that our proteins are functional. Moreover, only the proteins found in the clade of BclA and BacA were capable of transporting NCR247, while the ExsE and *Mycobacterium* BacA transporters were not able to complement the defect of the of *S. meliloti bacA* mutant. Furthermore, the ability of all non-symbiotic BclA transporters to transport NCR247 does not mean that they are all able to transport other NCR peptides. It's mean that NCR247 do not require a specific BacA/BclA to be imported, consistent with the fact that BacA/BclA are a broad antimicrobial peptide transporter. Nevertheless, other NCR peptides may require a host-specific BacA/BclA transporter. Indeed, these tested non-symbiotic BclA proteins were not able to complement the symbiotic defect of the *S. meliloti bacA* mutant *in planta*, suggesting that they were not able to import all the NCR peptides produced by *M. sativa.*

Therefore, we suggest that BacA and BclA were co-opted at least two times independently to cope with NCR peptides. A deep dN/dS selection analysis to assess whether a purifying or a positive selection occurs in the branches where BacA and BclA are acquired will be complementary to this analysis. Moreover, the reconstruction of the ancestral state of BacA and BclA could give new insights about their evolution.

D. Are NCR peptides and BacA transporters co-evolving independently?

The symbiotic relationships between plants and bacteria are intricate and dynamic, with bacteria often adapting to plants due to differences in genomic complexity and evolutionary timelines. Indeed, different studies showed a high variability in the transcriptomes of different accessions of *M. truncatula* with the effects being more dependent on the host genotype than on the bacterial symbiont genotype (Mergaert, 2018).

Even though we found that NCR247 is transported by different BclA transporters belonging to bacterial species that are not involved in symbiosis, it is possible that other NCR peptides are transported only by their symbiont BacA transporter. Indeed, NCR247 is a small monotypic peptide that does not share any sequence or structure similarity with other NCR peptides, and thus, does not represent the NCR peptides. Therefore, we believe that even though some NCR peptides are transported by different BacA or BclA transporters, there are some NCR peptides that are transported only by a compatible BacA transporter, which suggest an adaptation of BacA transporters to their host NCR peptides. For instance, it has been suggested that the NCR peptides of *Aeschynomene* legumes are not transported by *S. meliloti* BacA (Guefrachi et al., 2015). It is possible that *S. meliloti* BacA is not able to transport type2 NCRs with 8 cysteines. Moreover, the ability of *S. meliloti* BacA to cope with highly cationic NCR peptides of *M. sativa*, while *R. leguminosarum* and *S. fredii* can support the symbiosis with *M. officinalis* with less cationic NCRs but not with *M. sativa* (Huang et al., 2022).

In the other hand, we suggested above that the diversity of the numbers, sequences and structures of NCR peptides among legume species that induce TBD is related to the ability of BacA transporter of their symbionts to transport these peptides. Thus, it is possible that the rhizobial symbionts selects a specific NCR peptide according to the ability of their BacA to transport them. Therefore, we believe that the diversity of NCR peptides at different scales (between legume clades, legume species, and species accessions) is the result of adaption to different molecular actors of the rhizobial symbionts, including their transporters (BacA or BclA).

In summary, we suggest that a convergent coevolution occurs between NCR peptides and BacA transporters. However, whether specific BacA and BclA transporters are coevolving with specific NCR peptides remains to be elucidated by molecular biology guided by these structure-based evolutionary studies (**Figure 37**). In one hand, *I. argentea* NCR, which is different from

M. sativa (host of *S. meliloti*) NCRs, induced ploidy in *S. meliloti* suggests that the adaptability of BacA and BclA to cope with NCR peptides is not host specific. On the other hand, *S. meliloti* BacA are not able to transport *Aeschynomene* NCRs, requiring a host-specific BclA transport. These observations suggests as mentioned above that some NCR peptides are transported by different BacA/BclA transporters and do not require a host-specific transporter, while other NCR peptides (probably coevolving with their host-specific BacA) requires a host-specific BacA/BaclA transporter. Testing our identified NCR peptides with different BacA and BclA, including their symbionts BacA/BclA, and comparing between NCR-BacA (symbionts) and NCR-BacA (not symbionts) will allow us to gain more insights about the specificity of this coevolution. Moreover, generating different BacA alleles of the same symbiont will allow us to find specific positions that coevolve with NCR peptides. However, it is not evident to so the same with NCR peptides because they are short peptides, and a simple point mutation can alter their function. Furthermore, it is possible that other bacterial molecular actors could also coevolve with NCR peptides, such as the YejABEF transporter, the other transporter of NCR peptides and cell cycle regulators that interact with NCR peptides.

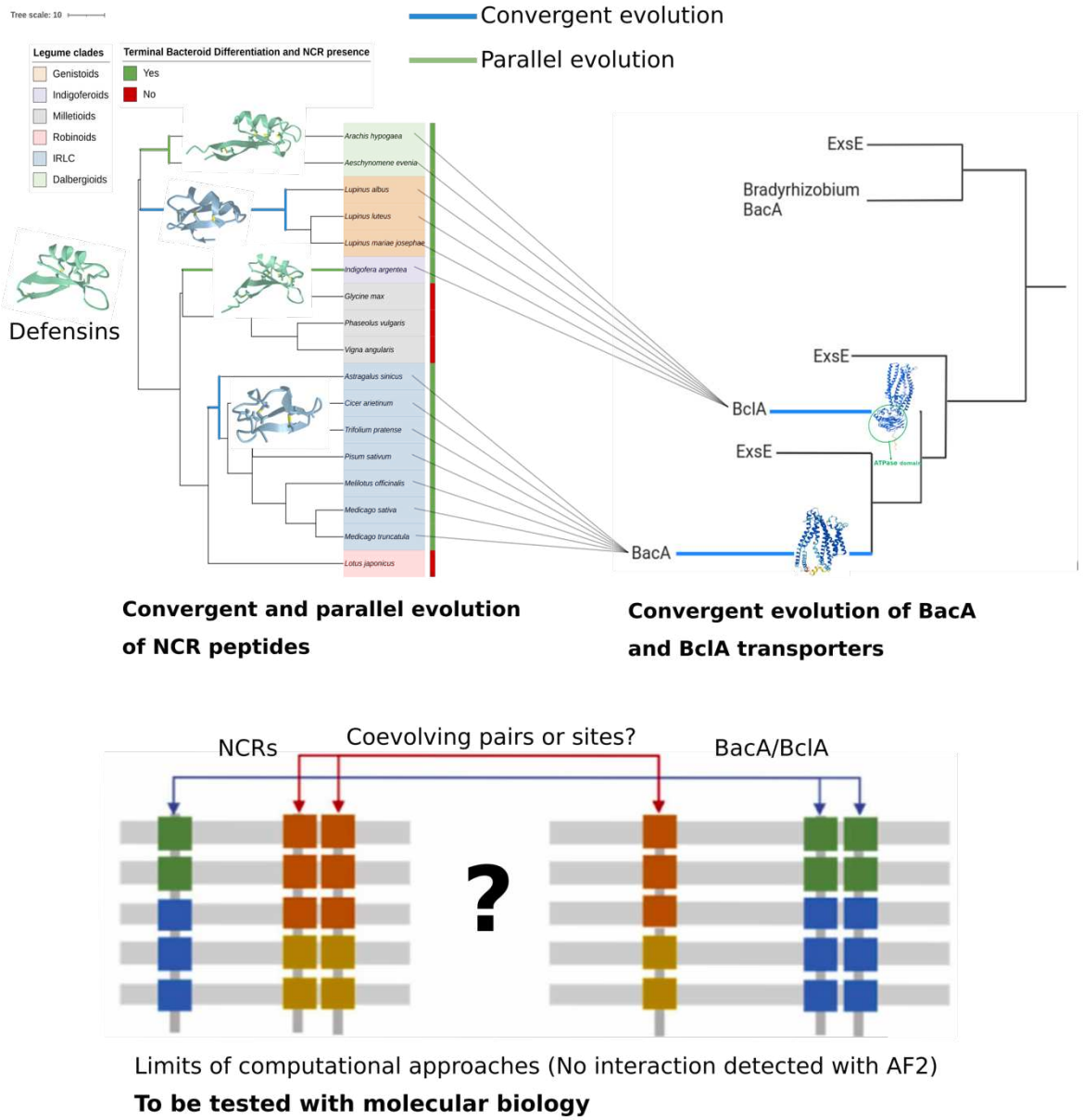
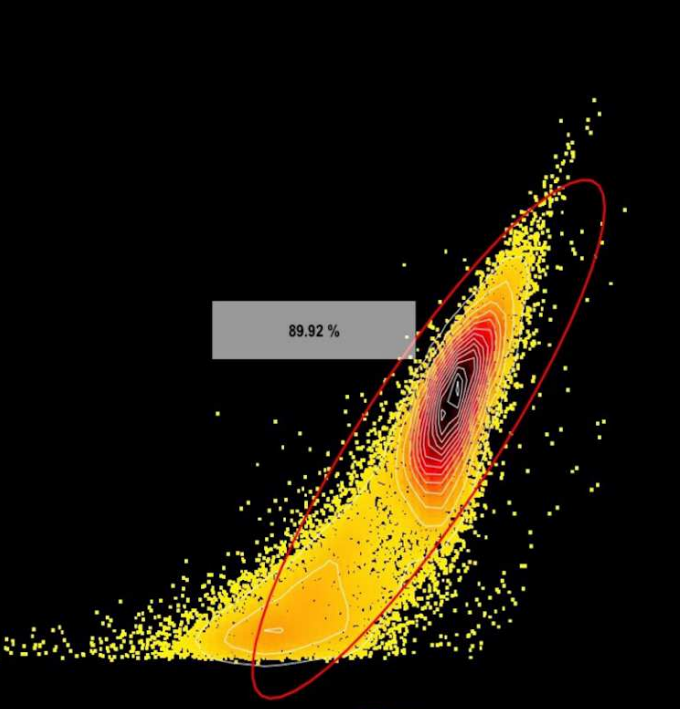


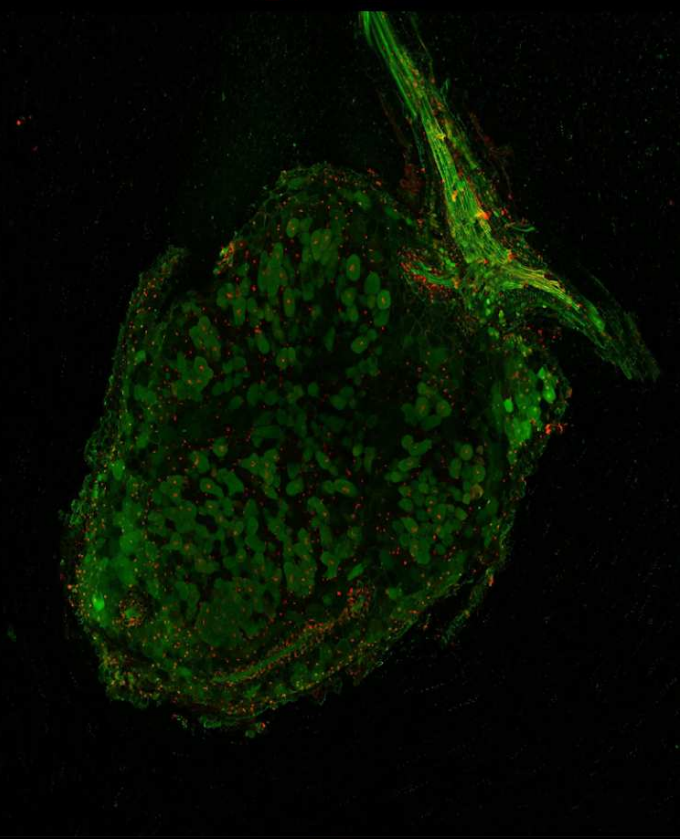
Figure 37 Convergent and parallel evolution of NCR peptides and the independent recruitment of BacA transporters to cope with NCR peptides. Do a convergent coevolution drive the evolution of NCR peptides and BacA transporters?

Conclusion

With the increasing availability of plant and bacteria genomes and the emergence of new sequencing technologies, the detection and classification of new biological sequences became fast and affordable. Moreover, the advancements in structural bioinformatics open up new opportunities to infer the molecular evolution of these sequences and test if their evolutionary history reflects the evolution of specific phenotypes in their species. Here, the combination of sequence-based and structural bioinformatics with functional experiments allowed us to decipher the evolutionary history of NCR peptides and show the complexity of evolution. While the phenotype of TBD has evolved independently multiple times to enhance host fitness benefits, the evolution of NCR peptides was more complex and showed both convergent, parallel, and divergent evolution. Moreover, we provide an excellent model to further study the convergent coevolution between NCR peptides and BacA transporters.

In addition to the evolutionary side, this study opens up new perspectives to study the diversity of NCR peptides at the functional level to gain more insights about the molecular actors of an advantageous trait that evolved in legumes, the TBD. For instance, are putative NCR-like peptides in species that do not induce TBD functional? Do all NCR peptides have the same promoter regions and AHL binding motifs?





References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Adhikari, B., Bhattacharya, D., Cao, R., & Cheng, J. (2017). Assessing Predicted Contacts for Building Protein Three-Dimensional Models. In Y. Zhou, A. Kloczkowski, E. Faraggi, & Y. Yang (Eds.), *Prediction of Protein Secondary Structure* (pp. 115–126). Springer. https://doi.org/10.1007/978-1-4939-6406-2_9

Agrawal, A. A., & Zhang, X. (2021). The evolution of coevolution in the study of species interactions. *Evolution*, *75*(7), 1594–1606. https://doi.org/10.1111/evo.14293

Alunni, B., & Gourion, B. (2016). Terminal bacteroid differentiation in the legume−rhizobium symbiosis: Nodule-specific cysteine-rich peptides and beyond. *New Phytologist*, *211*(2), 411–417. https://doi.org/10.1111/nph.14025

Alunni, B., Kevei, Z., Redondo-Nieto, M., Kondorosi, A., Mergaert, P., & Kondorosi, E. (2007). Genomic Organization and Evolutionary Insights on GRP and NCR Genes, Two Large Nodule-Specific Gene Families in Medicago truncatula. *Molecular Plant-Microbe Interactions®*, *20*(9), 1138–1148. https://doi.org/10.1094/MPMI-20-9-1138

Amarowicz, R. (2020). Legume Seeds as an Important Component of Human Diet. *Foods*, *9*(12), Article 12. https://doi.org/10.3390/foods9121812

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Antunes, L. C., Poppleton, D., Klingl, A., Criscuolo, A., Dupuy, B., Brochier-Armanet, C., Beloin, C., & Gribaldo, S. (2016). Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife*, *5*, e14589. https://doi.org/10.7554/eLife.14589

Arashida, H., Odake, H., Sugawara, M., Noda, R., Kakizaki, K., Ohkubo, S., Mitsui, H., Sato, S., & Minamisawa, K. (2022). Evolution of rhizobial symbiosis islands through insertion sequence-mediated deletion and duplication. *The ISME Journal*, *16*(1), 112–121. https://doi.org/10.1038/s41396-021-01035-4

Arbuckle, K., Bennett, C. M., & Speed, M. P. (2014). A simple measure of the strength of convergent evolution. *Methods in Ecology and Evolution*, *5*(7), 685–693. https://doi.org/10.1111/2041-210X.12195

Arbuckle, K., Rodríguez de la Vega, R. C., & Casewell, N. R. (2017). Coevolution takes the sting out of it: Evolutionary biology and mechanisms of toxin resistance in animals. *Toxicon*, *140*, 118–131. https://doi.org/10.1016/j.toxicon.2017.10.026

Archibald, J. M. (2015). Endosymbiosis and Eukaryotic Cell Evolution. *Current Biology*, *25*(19), R911–R921. https://doi.org/10.1016/j.cub.2015.07.055

Ardissone, S., Kobayashi, H., Kambara, K., Rummel, C., Noel, K. D., Walker, G. C., Broughton, W. J., & Deakin, W. J. (2011). Role of BacA in Lipopolysaccharide Synthesis, Peptide Transport, and Nodulation by Rhizobium sp. Strain NGR234. *Journal of Bacteriology*,

193(9), 2218–2228. https://doi.org/10.1128/JB.01260-10

Arnold, M. F. F., Haag, A. F., Capewell, S., Boshoff, H. I., James, E. K., McDonald, R., Mair, I., Mitchell, A. M., Kerscher, B., Mitchell, T. J., Mergaert, P., Barry, C. E., Scocchi, M., Zanda, M., Campopiano, D. J., & Ferguson, G. P. (2013). Partial Complementation of Sinorhizobium meliloti bacA Mutant Phenotypes by the Mycobacterium tuberculosis BacA Protein. *Journal of Bacteriology*, *195*(2), 389–398. https://doi.org/10.1128/jb.01445-12

Audain, E., Ramos, Y., Hermjakob, H., Flower, D. R., & Perez-Riverol, Y. (2016). Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. *Bioinformatics*, *32*(6), 821–827. https://doi.org/10.1093/bioinformatics/btv674

Azani, N., Babineau, M., Bailey, C. D., Banks, H., Barbosa, A. R., Pinto, R. B., Boatwright, J. S., Borges, L. M., Brown, G. K., Bruneau, A., Candido, E., Cardoso, D., Chung, K.-F., Clark, R. P., Conceição, A. de S., Crisp, M., Cubas, P., Delgado-Salinas, A., Dexter, K. G., … Zimmerman, E. (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). *TAXON*, *66*(1), 44–77. https://doi.org/10.12705/661.3

Barrière, Q., Guefrachi, I., Gully, D., Lamouche, F., Pierre, O., Fardoux, J., Chaintreuil, C., Alunni, B., Timchenko, T., Giraud, E., & Mergaert, P. (2017). Integrated roles of BclA and DD-carboxypeptidase 1 in Bradyrhizobium differentiation within NCR-producing and NCR-lacking root nodules. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-08830-0

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L. M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., & Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. *Nature*, *622*(7983), 637–645. https://doi.org/10.1038/s41586-023-06510-w

Bastolla, U., Porto, M., Eduardo Roman, M. H., & Vendruscolo, M. H. (2003). Connectivity of Neutral Networks, Overdispersion, and Structural Conservation in Protein Evolution. *Journal of Molecular Evolution*, *56*(3), 243–254. https://doi.org/10.1007/s00239-002-2350-0

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., & Sonnhammer, E. L. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Research*, *27*(1), 260–262. https://doi.org/10.1093/nar/27.1.260

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*(suppl_1), D138–D141. https://doi.org/10.1093/nar/gkh121

Bellaire, B. H., Roop, R. M., & Cardelli, J. A. (2005). Opsonized Virulent Brucella abortus Replicates within Nonacidic, Endoplasmic Reticulum-Negative, LAMP-1-Positive Phagosomes in Human Monocytes. *Infection and Immunity*, *73*(6), 3702–3713. https://doi.org/10.1128/iai.73.6.3702-3713.2005

Berendzen, J., Brown, A. V., Cameron, C. T., Campbell, J. D., Cleary, A. M., Dash, S., Hokin, S., Huang, W., Kalberer, S. R., Nelson, R. T., Redsun, S., Weeks, N. T., Wilkey, A., Farmer, A. D., & Cannon, S. B. (2021). The legume information system and associated online genomic resources. *Legume Science*, *3*(3), e74. https://doi.org/10.1002/leg3.74

Bergman, B., Johansson, C., & Söderbäck, E. (1992). The Nostoc–Gunnera symbiosis. *New Phytologist*, *122*(3), 379–400. https://doi.org/10.1111/j.1469-8137.1992.tb00067.x

Betts, H. C., Puttick, M. N., Clark, J. W., Williams, T. A., Donoghue, P. C. J., & Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology & Evolution*, *2*(10), 1556–1562. https://doi.org/10.1038/s41559-018-0644-x

Bhattacharya, D., Yoon, H. S., & Hackett, J. D. (2004). Photosynthetic eukaryotes unite: Endosymbiosis connects the dots. *BioEssays*, *26*(1), 50–60. https://doi.org/10.1002/bies.10376

Blount, Z. D., Lenski, R. E., & Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science*, *362*(6415), eaam5979. https://doi.org/10.1126/science.aam5979

Bordin, N., Sillitoe, I., Nallapareddy, V., Rauer, C., Lam, S. D., Waman, V. P., Sen, N., Heinzinger, M., Littmann, M., Kim, S., Velankar, S., Steinegger, M., Rost, B., & Orengo, C. (2023). AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology*, *6*(1), 1–12. https://doi.org/10.1038/s42003-023-04488-9

Boukherissa, A., Antaya, K., Howe, G. W., Vega, R. C. R. de la, Shykoff, J. A., Alunni, B., & diCenzo, G. C. (2024). *Taxonomic distribution of SbmA/BacA and BacA-like antimicrobial peptide transporters suggests independent recruitment and convergent evolution in host-microbe interactions* (p. 2024.02.25.581009). bioRxiv. https://doi.org/10.1101/2024.02.25.581009

Bouznif, B., Boukherissa, A., Jaszczyszyn, Y., Mars, M., Timchenko, T., Shykoff, J. A., & Alunni, B. (2024). Complete and circularized genome sequences of five nitrogen-fixing Bradyrhizobium sp. Strains isolated from root nodules of peanut, Arachis hypogaea, cultivated in Tunisia. *Microbiology Resource Announcements*, *13*(6), e01078-23. https://doi.org/10.1128/mra.01078-23

Boyd, B. M., James, I., Johnson, K. P., Weiss, R. B., Bush, S. E., Clayton, D. H., & Dale, C. (2024). Stochasticity, determinism, and contingency shape genome evolution of endosymbiotic bacteria. *Nature Communications*, *15*(1), 4571. https://doi.org/10.1038/s41467-024-48784-2

Brandon, R. N. (1978). Adaptation and evolutionary theory. *Studies in History and Philosophy of Science Part A*, *9*(3), 181–206. https://doi.org/10.1016/0039-3681(78)90005-5

Brazhnik, P., & Tyson, J. J. (2006). Cell Cycle Control in Bacteria and Yeast: A Case of Convergent Evolution? *Cell Cycle*, *5*(5), 522–529. https://doi.org/10.4161/cc.5.5.2493

Brewin, N. J. (1991). Development of the Legume Root Nodule. *Annual Review of Cell and Developmental Biology*, *7*(Volume 7, 1991), 191–226. https://doi.org/10.1146/annurev.cb.07.110191.001203

Brockhurst, M. A., Rainey, P. B., & Buckling, A. (2004). The effect of spatial heterogeneity and parasites on the evolution of host diversity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1534), 107–111. https://doi.org/10.1098/rspb.2003.2556

Brown, R. F. (2016). *Investigating the evolutionary origins and cell biology of Negativicutes.*

[Phd, University of Warwick]. http://webcat.warwick.ac.uk/record=b3067395~S15

Buckling, A., & Rainey, P. B. (2002a). The role of parasites in sympatric and allopatric host diversification. *Nature*, *420*(6915), 496–499. https://doi.org/10.1038/nature01164

Buckling, A., & Rainey, P. B. . (2002b). Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1494), 931–936. https://doi.org/10.1098/rspb.2001.1945

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carmona, D., Fitzpatrick, C. R., & Johnson, M. T. J. (2015). Fifty years of co-evolution and beyond: Integrating co-evolution from molecules to species. *Molecular Ecology*, *24*(21), 5315–5329. https://doi.org/10.1111/mec.13389

Cerca, J. (2023). Understanding natural selection and similarity: Convergent, parallel and repeated evolution. *Molecular Ecology*, *32*(20), 5451–5462. https://doi.org/10.1111/mec.17132

Chaudhary, R., Burleigh, J. G., & Fernandez-Baca, D. (2012). Fast Local Search for Unrooted Robinson-Foulds Supertrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1004–1013. IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://doi.org/10.1109/TCBB.2012.47

Checcucci, A., diCenzo, G. C., Perrin, E., Bazzicalupo, M., & Mengoni, A. (2019). Chapter 3—Genomic Diversity and Evolution of Rhizobia. In S. Das & H. R. Dash (Eds.), *Microbial Diversity in the Genomic Era* (pp. 37–46). Academic Press. https://doi.org/10.1016/B978-0-12-814849-5.00003-4

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Chen, S., Rudra, B., & Gupta, R. S. (2021). Phylogenomics and molecular signatures support division of the order Neisseriales into emended families Neisseriaceae and Chromobacteriaceae and three new families Aquaspirillaceae fam. Nov., Chitinibacteraceae fam. Nov., and Leeiaceae fam. Nov. *Systematic and Applied Microbiology*, *44*(6), 126251. https://doi.org/10.1016/j.syapm.2021.126251

Cherkasov, N., Ibhadon, A. O., & Fitzpatrick, P. (2015). A review of the existing and alternative methods for greener nitrogen fixation. *Chemical Engineering and Processing: Process Intensification*, *90*, 24–33. https://doi.org/10.1016/j.cep.2015.02.004

Chirat, R., Moulton, D. E., & Goriely, A. (2013). Mechanical basis of morphogenesis and convergent evolution of spiny seashells. *Proceedings of the National Academy of Sciences*, *110*(15), 6015–6020. https://doi.org/10.1073/pnas.1220443110

Choi, I.-S., Cardoso, D., de Queiroz, L. P., de Lima, H. C., Lee, C., Ruhlman, T. A., Jansen, R.

K., & Wojciechowski, M. F. (2022). Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.823190

Codoñer, F. M., & Fares, M. A. (2008). Why Should We Care about Molecular Coevolution? *Evolutionary Bioinformatics*, *4*, 117693430800400003. https://doi.org/10.1177/117693430800400003

Collins, D. (1990). [Review of *Review of WONDERFUL LIFE. THE BURGESS SHALE AND THE NATURE OF HISTORY*, by S. J. Gould]. *Earth Sciences History*, *9*(2), 163–165.

Collins, T. J. (2007). ImageJ for Microscopy. *BioTechniques*, *43*(sup1), S25–S30. https://doi.org/10.2144/000112517

Couturier, C., Groß, S., Tesmar, A. von, Hoffmann, J., Deckarm, S., Fievet, A., Dubarry, N., Taillier, T., Pöverlein, C., Stump, H., Kurz, M., Toti, L., Richter, S. H., Schummer, D., Sizun, P., Hoffmann, M., Awal, R. P., Zaburannyi, N., Harmrolfs, K., … Renard, S. (2022). *Structure elucidation, biosynthesis, total synthesis and antibacterial in-vivo efficacy of myxobacterial Corramycin*. ChemRxiv. https://doi.org/10.26434/chemrxiv-2022-97gp2

Czernic, P., Gully, D., Cartieaux, F., Moulin, L., Guefrachi, I., Patrel, D., Pierre, O., Fardoux, J., Chaintreuil, C., Nguyen, P., Gressent, F., Da Silva, C., Poulain, J., Wincker, P., Rofidal, V., Hem, S., Barrière, Q., Arrighi, J.-F., Mergaert, P., & Giraud, E. (2015). Convergent Evolution of Endosymbiont Differentiation in Dalbergioid and Inverted Repeat-Lacking Clade Legumes Mediated by Nodule-Specific Cysteine-Rich Peptides. *Plant Physiology*, *169*(2), 1254–1265. https://doi.org/10.1104/pp.15.00584

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dart, P. J., & Mercer, F. V. (1966). Fine Structure of Bacteroids in Root Nodules of Vigna sinensis, Acacia longifolia, Viminaria juncea, and Lupinus angustifolius. *Journal of Bacteriology*, *91*(3), 1314–1319. https://doi.org/10.1128/jb.91.3.1314-1319.1966

Darwin, C. (1859). *On the origin of species: Facsimile of the first edition*.

Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: A data driven reference for analysis and visualisation of transcriptomes. *Genome Biology*, *18*(1), 148. https://doi.org/10.1186/s13059-017-1284-1

Davidson, N. M., & Oshlack, A. (2014). Corset: Enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biology*, *15*(7), 410. https://doi.org/10.1186/s13059-014-0410-6

de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, *14*(4), 249–261. https://doi.org/10.1038/nrg3414

de la Maza LM. (2020). Introduction to Enterobacterales. In *Color Atlas of Medical Bacteriology* (pp. 91–102). John Wiley & Sons, Ltd. https://doi.org/10.1128/9781683671077.ch10

De Lajudie, P., LAURENT-FULELE, E., WILLEMS, A., TOREK, U., COOPMAN, R., COLLINS, M. D., KERSTERS, K., DREYFUS, B., & GILLIS, M. (1998). Allorhizobium undicola gen. Nov., sp. Nov., nitrogen-fixing bacteria that efficiently nodulate Neptunia natans in Senegal. *International Journal of Systematic and Evolutionary Microbiology*, *48*(4), 1277–1290. https://doi.org/10.1099/00207713-48-4-1277

Delaux, P.-M., Radhakrishnan, G., & Oldroyd, G. (2015). Tracing the evolutionary path to nitrogen-fixing crops. *Current Opinion in Plant Biology*, *26*, 95–99. https://doi.org/10.1016/j.pbi.2015.06.003

Dendene, S., Frascella, A., Nicoud, Q., Timchenko, T., Mergaert, P., Alunni, B., & Biondi, E. G. (2022). Cell Cycle and Terminal Differentiation in Sinorhizobium meliloti. In E. Biondi (Ed.), *Cell Cycle Regulation and Development in Alphaproteobacteria* (pp. 221–244). Springer International Publishing. https://doi.org/10.1007/978-3-030-90621-4_8

Dendene, S., Xue, S., Nicoud, Q., Valette, O., Frascella, A., Bonnardel, A., Bars, R. L., Bourge, M., Mergaert, P., Brilli, M., Alunni, B., & Biondi, E. G. (2023). *Sinorhizobium meliloti FcrX coordinates cell cycle and division during free-living growth and symbiosis* (p. 2023.03.13.532326). bioRxiv. https://doi.org/10.1101/2023.03.13.532326

diCenzo, G. C., Mengoni, A., & Perrin, E. (2019). Chromids Aid Genome Expansion and Functional Diversification in the Family Burkholderiaceae. *Molecular Biology and Evolution*, *36*(3), 562–574. https://doi.org/10.1093/molbev/msy248

diCenzo, G. C., Tesi, M., Pfau, T., Mengoni, A., & Fondi, M. (2020). Genome-scale metabolic reconstruction of the symbiosis between a leguminous plant and a nitrogen-fixing bacterium. *Nature Communications*, *11*(1), 1–11.

diCenzo, G. C., Yang, Y., Young, J. P. W., & Kuzmanović, N. (2023). *Refining the taxonomy of the order Hyphomicrobiales (Rhizobiales) based on whole genome comparisons of over 130 genus type strains* (p. 2023.11.15.567303). bioRxiv. https://doi.org/10.1101/2023.11.15.567303

diCenzo, G. C., Zamani, M., Ludwig, H. N., & Finan, T. M. (2017). Heterologous Complementation Reveals a Specialized Activity for BacA in the Medicago–Sinorhizobium meliloti Symbiosis. *Molecular Plant-Microbe Interactions®*, *30*(4), 312–324. https://doi.org/10.1094/MPMI-02-17-0030-R

Dinkins, R. D., Hancock, J. A., Bickhart, D. M., Sullivan, M. L., & Zhu, H. (2022). Expression and Variation of the Genes Involved in Rhizobium Nodulation in Red Clover. *Plants*, *11*(21), Article 21. https://doi.org/10.3390/plants11212888

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dobritsa, A. P., & Samadpour, M. (2016). Transfer of eleven species of the genus Burkholderia to the genus Paraburkholderia and proposal of Caballeronia gen. Nov. To accommodate twelve species of the genera Burkholderia and Paraburkholderia. *International Journal of Systematic and Evolutionary Microbiology*, *66*(8), 2836–2846. https://doi.org/10.1099/ijsem.0.001065

Domenech, P., Kobayashi, H., LeVier, K., Walker, G. C., & Barry, C. E. (2009). BacA, an ABC Transporter Involved in Maintenance of Chronic Murine Infections with Mycobacterium tuberculosis. *Journal of Bacteriology*, *191*(2), 477–485. https://doi.org/10.1128/jb.01132-08

Doolittle, R. F. (1994). Convergent evolution: The need to be explicit. *Trends in Biochemical Sciences*, *19*(1), 15–18. https://doi.org/10.1016/0968-0004(94)90167-8

Douam, F., Fusil, F., Enguehard, M., Dib, L., Nadalin, F., Schwaller, L., Hrebikova, G., Mancip, J., Mailly, L., Montserret, R., Ding, Q., Maisse, C., Carlot, E., Xu, K., Verhoeyen, E., Baumert, T. F., Ploss, A., Carbone, A., Cosset, F.-L., & Lavillette, D. (2018). A protein coevolution method uncovers critical features of the Hepatitis C Virus fusion mechanism. *PLOS Pathogens*, *14*(3), e1006908. https://doi.org/10.1371/journal.ppat.1006908

Dovrat, G., & Sheffer, E. (2019). Symbiotic dinitrogen fixation is seasonal and strongly regulated in water-limited environments. *New Phytologist*, *221*(4), 1866–1877. https://doi.org/10.1111/nph.15526

Downie, J. A., & Kondorosi, E. (2021). Why Should Nodule Cysteine-Rich (NCR) Peptides Be Absent From Nodules of Some Groups of Legumes but Essential for Symbiotic N-Fixation in Others? *Frontiers in Agronomy*, *3*. https://www.frontiersin.org/articles/10.3389/fagro.2021.654576

Durgo, H., Klement, E., Hunyadi-Gulyas, E., Szucs, A., Kereszt, A., Medzihradszky, K. F., & Kondorosi, E. (2015). Identification of nodule-specific cysteine-rich plant peptides in endosymbiotic bacteria. *PROTEOMICS*, *15*(13), 2291–2295. https://doi.org/10.1002/pmic.201400385

Eddy, S. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*. https://doi.org/10.1142/9781848165632_0019

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and Plants: A Study in Coevolution. *Evolution*, *18*(4), 586–608. https://doi.org/10.2307/2406212

Ekseth, O. K., Kuiper, M., & Mironov, V. (2014). orthAgogue: An agile tool for the rapid prediction of orthology relations. *Bioinformatics*, *30*(5), 734–736. https://doi.org/10.1093/bioinformatics/btt582

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. https://doi.org/10.1093/nar/30.7.1575

Erdős, G., & Dosztányi, Z. (2023). Chapter 7—Prediction of protein structure and intrinsic disorder in the era of deep learning. In M. N. Gupta & V. N. Uversky (Eds.), *Structure and Intrinsic Disorder in Enzymology* (pp. 199–224). Academic Press. https://doi.org/10.1016/B978-0-323-99533-7.00007-8

Eswarappa, S. M., Panguluri, K. K., Hensel, M., & Chakravortty, D. (2008). The yejABEF operon of Salmonella confers resistance to antimicrobial peptides and contributes to its virulence. *Microbiology*, *154*(2), 666–678. https://doi.org/10.1099/mic.0.2007/011114-0

Farkas, A., Maróti, G., Dürgő, H., Györgypál, Z., Lima, R. M., Medzihradszky, K. F., Kereszt, A., Mergaert, P., & Kondorosi, É. (2014). Medicago truncatula symbiotic peptide NCR247 contributes to bacteroid differentiation through multiple mechanisms. *Proceedings of the National Academy of Sciences*, *111*(14), 5183–5188. https://doi.org/10.1073/pnas.1404169111

Feigin, C. Y., Newton, A. H., Doronina, L., Schmitz, J., Hipsley, C. A., Mitchell, K. J., Gower, G., Llamas, B., Soubrier, J., Heider, T. N., Menzies, B. R., Cooper, A., O'Neill, R. J., & Pask, A. J. (2018). Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. *Nature Ecology & Evolution*, *2*(1), 182–192. https://doi.org/10.1038/s41559-017-0417-y

Ferguson, G. P., Datta, A., Baumgartner, J., Roop, R. M., Carlson, R. W., & Walker, G. C. (2004). Similarity to peroxisomal-membrane protein family reveals that Sinorhizobium and Brucella BacA affect lipid-A fatty acids. *Proceedings of the National Academy of Sciences*, *101*(14), 5012–5017. https://doi.org/10.1073/pnas.0307137101

Ferguson, G. P., Datta, A., Carlson, R. W., & Walker, G. C. (2005). Importance of unusually modified lipid A in Sinorhizobium stress resistance and legume symbiosis. *Molecular Microbiology*, *56*(1), 68–80. https://doi.org/10.1111/j.1365-2958.2005.04536.x

Ferguson, G. P., Roop, R. M., & Walker, G. C. (2002). Deficiency of a Sinorhizobium meliloti bacA Mutant in Alfalfa Symbiosis Correlates with Alteration of the Cell Envelope. *Journal of Bacteriology*, *184*(20), 5625–5632. https://doi.org/10.1128/jb.184.20.5625-5632.2002

Fernández-Pascual, M., Pueyo, J. J., Felipe, M. R. de, Golvano, M. P., & Lucas, M. M. (2007). *Singular Features of the Bradyrhizobium-Lupinus Symbiosis*. https://digital.csic.es/handle/10261/12646

Finan, T. M., Kunkel, B., De Vos, G. F., & Signer, E. R. (1986). Second symbiotic megaplasmid in Rhizobium meliloti carrying exopolysaccharide and thiamine synthesis genes. *Journal of Bacteriology*, *167*(1), 66–72. https://doi.org/10.1128/jb.167.1.66-72.1986

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, *42*(D1), D222–D230. https://doi.org/10.1093/nar/gkt1223

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., & Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*(W1), W30–W38. https://doi.org/10.1093/nar/gkv397

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl_2), W29–W37. https://doi.org/10.1093/nar/gkr367

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C.,

Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279-285. https://doi.org/10.1093/nar/gkv1344

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., & Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, *38*(suppl_1), D211–D222. https://doi.org/10.1093/nar/gkp985

Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, *20*(4), 406–416. https://doi.org/10.1093/sysbio/20.4.406

Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., Khan, Z., Kovar, C., Lee, S. L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J., Raney, B. J., … Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, *47*(3), 272–275. https://doi.org/10.1038/ng.3198

Forde, S. E., Thompson, J. N., Holt, R. D., & Bohannan, B. J. M. (2008). COEVOLUTION DRIVES TEMPORAL CHANGES IN FITNESS AND DIVERSITY ACROSS ENVIRONMENTS IN A BACTERIA–BACTERIOPHAGE INTERACTION. *Evolution*, *62*(8), 1830–1839. https://doi.org/10.1111/j.1558-5646.2008.00411.x

Frickey, T., & Lupas, A. (2004). CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, *20*(18), 3702–3704. https://doi.org/10.1093/bioinformatics/bth444

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *304B*(1), 64–74. https://doi.org/10.1002/jez.b.21026

Gage, D. J. (2004). Infection and Invasion of Roots by Symbiotic, Nitrogen-Fixing Rhizobia during Nodulation of Temperate Legumes. *Microbiology and Molecular Biology Reviews*, *68*(2), 280–300. https://doi.org/10.1128/mmbr.68.2.280-300.2004

Garrido-Oter, R., Nakano, R. T., Dombrowski, N., Ma, K.-W., McHardy, A. C., & Schulze-Lefert, P. (2018). Modular Traits of the Rhizobiales Root Microbiota and Their Evolutionary Relationship with Symbiotic Rhizobia. *Cell Host & Microbe*, *24*(1), 155-167.e5. https://doi.org/10.1016/j.chom.2018.06.006

Garrity, G. M., Bell, J. A., & Lilburn, T. (2005). Pasteurellalesord. Nov. In D. J. Brenner, N. R. Krieg, J. T. Staley, G. M. Garrity, D. R. Boone, P. De Vos, M. Goodfellow, F. A. Rainey, & K.-H. Schleifer (Eds.), *Bergey's Manual® of Systematic Bacteriology: Volume Two The Proteobacteria Part B The Gammaproteobacteria* (pp. 850–912). Springer US.

https://doi.org/10.1007/0-387-28022-7_14

Gehlot, H. S., Tak, N., Kaushik, M., Mitra, S., Chen, W.-M., Poweleit, N., Panwar, D., Poonar, N., Parihar, R., Tak, A., Sankhla, I. S., Ojha, A., Rao, S. R., Simon, M. F., Reis Junior, F. B. dos, Perigolo, N., Tripathi, A. K., Sprent, J. I., Young, J. P. W., … Gyaneshwar, P. (2013). An invasive Mimosa in India does not adopt the symbionts of its native relatives. *Annals of Botany*, *112*(1), 179–196. https://doi.org/10.1093/aob/mct112

Ghilarov, D., Inaba-Inoue, S., Stepien, P., Qu, F., Michalczyk, E., Pakosz, Z., Nomura, N., Ogasawara, S., Walker, G. C., Rebuffat, S., Iwata, S., Heddle, J. G., & Beis, K. (2021). Molecular mechanism of SbmA, a promiscuous transporter exploited by antimicrobial peptides. *Science Advances*, *7*(37), eabj5363. https://doi.org/10.1126/sciadv.abj5363

Glazebrook, J., Ichige, A., & Walker, G. C. (1993). A Rhizobium meliloti homolog of the Escherichia coli peptide-antibiotic transport protein SbmA is essential for bacteroid development. *Genes & Development*, *7*(8), 1485–1497. https://doi.org/10.1101/gad.7.8.1485

Gompel, N., & Prud'homme, B. (2009). The causes of repeated genetic evolution. *Developmental Biology*, *332*(1), 36–47. https://doi.org/10.1016/j.ydbio.2009.04.040

Gopinath, K., Venclovas, Č., Ioerger, T. R., Sacchettini, J. C., McKinney, J. D., Mizrahi, V., & Warner, D. F. (2013). A vitamin B12 transporter in Mycobacterium tuberculosis. *Open Biology*, *3*(2), 120175. https://doi.org/10.1098/rsob.120175

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), Article 7. https://doi.org/10.1038/nbt.1883

Guefrachi, I., Nagymihaly, M., Pislariu, C. I., Van de Velde, W., Ratet, P., Mars, M., Udvardi, M. K., Kondorosi, E., Mergaert, P., & Alunni, B. (2014). Extreme specificity of NCR gene expression in Medicago truncatula. *BMC Genomics*, *15*(1), 712. https://doi.org/10.1186/1471-2164-15-712

Guefrachi, I., Pierre, O., Timchenko, T., Alunni, B., Barrière, Q., Czernic, P., Villaécija-Aguilar, J.-A., Verly, C., Bourge, M., Fardoux, J., Mars, M., Kondorosi, E., Giraud, E., & Mergaert, P. (2015). Bradyrhizobium BclA Is a Peptide Transporter Required for Bacterial Differentiation in Symbiosis with Aeschynomene Legumes. *Molecular Plant-Microbe Interactions®*, *28*(11), 1155–1166. https://doi.org/10.1094/MPMI-04-15-0094-R

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Haag, A. F., Arnold, M. F. F., Myka, K. K., Kerscher, B., Dall'Angelo, S., Zanda, M., Mergaert, P., & Ferguson, G. P. (2013). Molecular insights into bacteroid development during Rhizobium–legume symbiosis. *FEMS Microbiology Reviews*, *37*(3), 364–383. https://doi.org/10.1111/1574-6976.12003

Haag, A. F., Baloban, M., Sani, M., Kerscher, B., Pierre, O., Farkas, A., Longhi, R., Boncompagni, E., Hérouart, D., Dall'Angelo, S., Kondorosi, E., Zanda, M., Mergaert, P., & Ferguson, G. P. (2011). Protection of Sinorhizobium against Host Cysteine-Rich Antimicrobial Peptides Is Critical for Symbiosis. *PLOS Biology*, *9*(10), e1001169. https://doi.org/10.1371/journal.pbio.1001169

Haag, A. F., & Mergaert, P. (2020). Terminal bacteroid differentiation in the Medicago–Rhizobium interaction – a tug of war between plant and bacteria. In *The Model Legume Medicago truncatula* (pp. 600–616). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119409144.ch75

Haber, F., & van Oordt, G. (1905). Über die Bildung von Ammoniak den Elementen. *Zeitschrift für anorganische Chemie*, *44*(1), 341–378. https://doi.org/10.1002/zaac.19050440122

Han, Z., Liu, Y., Deng, X., Liu, D., Liu, Y., Hu, Y., & Yan, Y. (2019). Genome-wide identification and expression analysis of expansin gene family in common wheat (Triticum aestivum L.). *BMC Genomics*, *20*(1), 101. https://doi.org/10.1186/s12864-019-5455-1

Harder, L. D., & Johnson, S. D. (2009). Darwin's beautiful contrivances: Evolutionary and functional evidence for floral adaptation. *New Phytologist*, *183*(3), 530–545. https://doi.org/10.1111/j.1469-8137.2009.02914.x

Heckmann, A. B., Lombardo, F., Miwa, H., Perry, J. A., Bunnewell, S., Parniske, M., Wang, T. L., & Downie, J. A. (2006). Lotus japonicus Nodulation Requires Two GRAS Domain Regulators, One of Which Is Functionally Conserved in a Non-Legume. *Plant Physiology*, *142*(4), 1739–1750. https://doi.org/10.1104/pp.106.089508

Heidstra, R., Geurts, R., Franssen, H., Spaink, H. P., van Kammen, A., & Bisseling, T. (1994). Root Hair Deformation Activity of Nodulation Factors and Their Fate on Vicia sativa. *Plant Physiology*, *105*(3), 787–797. https://doi.org/10.1104/pp.105.3.787

Hirsch, A. M. (1992). Developmental biology of legume nodulation. *New Phytologist*, *122*(2), 211–237. https://doi.org/10.1111/j.1469-8137.1992.tb04227.x

Hirsch, A. M., Lum, M. R., & Downie, J. A. (2001). What Makes the Rhizobia-Legume Symbiosis So Special? *Plant Physiology*, *127*(4), 1484–1492. https://doi.org/10.1104/pp.010866

Hodges, S. A. (1997). Floral Nectar Spurs and Diversification. *International Journal of Plant Sciences*, *158*(S6), S81–S88. https://doi.org/10.1086/297508

Holm, L. (2022). Dali server: Structural unification of protein families. *Nucleic Acids Research*, *50*(W1), W210–W215. https://doi.org/10.1093/nar/gkac387

Horváth, B., Domonkos, Á., Kereszt, A., Szűcs, A., Ábrahám, E., Ayaydin, F., Bóka, K., Chen, Y., Chen, R., Murray, J. D., Udvardi, M. K., Kondorosi, É., & Kaló, P. (2015). Loss of the nodule-specific cysteine rich peptide, NCR169, abolishes symbiotic nitrogen fixation in the Medicago truncatula dnf7 mutant. *Proceedings of the National Academy of Sciences*, *112*(49),

15232–15237. https://doi.org/10.1073/pnas.1500777112

Horváth, B., Güngör, B., Tóth, M., Domonkos, Á., Ayaydin, F., Saifi, F., Chen, Y., Biró, J. B., Bourge, M., Szabó, Z., Tóth, Z., Chen, R., & Kaló, P. (2023). *The Medicago truncatula nodule-specific cysteine-rich peptides, NCR343 and NCR-new35 are required for the maintenance of rhizobia in nitrogen-fixing nodules* (p. 2023.01.23.523609). bioRxiv. https://doi.org/10.1101/2023.01.23.523609

Howarth, R. W., Marino, R., Lane, J., & Cole, J. J. (1988). Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 1. Rates and importance1. *Limnology and Oceanography*, *33*(4part2), 669–687. https://doi.org/10.4319/lo.1988.33.4part2.0669

Howe, K., Bateman, A., & Durbin, R. (2002). QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, *18*(11), 1546–1547. https://doi.org/10.1093/bioinformatics/18.11.1546

Huang, R., Snedden, W. A., & diCenzo, G. C. (2022). Reference nodule transcriptomes for Melilotus officinalis and Medicago sativa cv. Algonquin. *Plant Direct*, *6*(6), e408. https://doi.org/10.1002/pld3.408

Ichige, A., & Walker, G. C. (1997). Genetic analysis of the Rhizobium meliloti bacA gene: Functional interchangeability with the Escherichia coli sbmA gene and phenotypes of mutants. *Journal of Bacteriology*, *179*(1), 209–216. https://doi.org/10.1128/jb.179.1.209-216.1997

Jang, S., Mergaert, P., Ohbayashi, T., Ishigami, K., Shigenobu, S., Itoh, H., & Kikuchi, Y. (2021). Dual oxidase enables insect gut symbiosis by mediating respiratory network formation. *Proceedings of the National Academy of Sciences*, *118*(10), e2020922118. https://doi.org/10.1073/pnas.2020922118

Jensen, H. L., & Jensen, H. L. (1942). Nitrogen fixation in leguminous plants. I. General characters of root-nodule bacteria isolated from species of Medicago and Trifolium in Australia. *Proceedings of the Linnean Society of New South Wales*, *67*, 98--108.

Jetten, L., & van Iersel, L. (2018). Nonbinary Tree-Based Phylogenetic Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(1), 205–217. IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://doi.org/10.1109/TCBB.2016.2615918

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*, 431. https://doi.org/10.1186/1471-2105-11-431

Jones, K. M., Kobayashi, H., Davies, B. W., Taga, M. E., & Walker, G. C. (2007). How rhizobial symbionts invade plants: The Sinorhizobium–Medicago model. *Nature Reviews Microbiology*, *5*(8), 619–633. https://doi.org/10.1038/nrmicro1705

Jones, R. M., & Neish, A. S. (2011). Recognition of bacterial pathogens and mucosal immunity. *Cellular Microbiology*, *13*(5), 670–676. https://doi.org/10.1111/j.1462-5822.2011.01579.x

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021).

Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kant, C., Pradhan, S., & Bhatia, S. (2016). Dissecting the Root Nodule Transcriptome of Chickpea (Cicer arietinum L.). *PLOS ONE*, *11*(6), e0157908. https://doi.org/10.1371/journal.pone.0157908

Karunakaran, R., Haag, A. F., East, A. K., Ramachandran, V. K., Prell, J., James, E. K., Scocchi, M., Ferguson, G. P., & Poole, P. S. (2010). BacA Is Essential for Bacteroid Development in Nodules of Galegoid, but not Phaseoloid, Legumes. *Journal of Bacteriology*, *192*(11), 2920–2928. https://doi.org/10.1128/JB.00020-10

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kazmierczak, T., Nagymihály, M., Lamouche, F., Barrière, Q., Guefrachi, I., Alunni, B., Ouadghiri, M., Ibijbijen, J., Kondorosi, É., Mergaert, P., & Gruber, V. (2017). Specific Host-Responsive Associations Between Medicago truncatula Accessions and Sinorhizobium Strains. *Molecular Plant-Microbe Interactions®*, *30*(5), 399–409. https://doi.org/10.1094/MPMI-01-17-0009-R

Kazmierczak, T., Yang, L., Boncompagni, E., Meilhoc, E., Frugier, F., Frendo, P., Bruand, C., Gruber, V., & Brouquisse, R. (2020). Chapter Seven - Legume nodule senescence: A coordinated death mechanism between bacteria and plant cells. In P. Frendo, F. Frugier, & C. Masson-Boivin (Eds.), *Advances in Botanical Research* (Vol. 94, pp. 181–212). Academic Press. https://doi.org/10.1016/bs.abr.2019.09.013

Keller, J., Imperial, J., Ruiz-Argüeso, T., Privet, K., Lima, O., Michon-Coudouel, S., Biget, M., Salmon, A., Aïnouche, A., & Cabello-Hurtado, F. (2018). RNA sequencing and analysis of three Lupinus nodulomes provide new insights into specific host-symbiont relationships with compatible and incompatible Bradyrhizobium strains. *Plant Science*, *266*, 102–116. https://doi.org/10.1016/j.plantsci.2017.10.015

Kereszt, A., Mergaert, P., Montiel, J., Endre, G., & Kondorosi, É. (2018). Impact of Plant Peptides on Symbiotic Nodule Development and Functioning. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01026

Kim, M., Chen, Y., Xi, J., Waters, C., Chen, R., & Wang, D. (2015). An antimicrobial peptide essential for bacterial survival in the nitrogen-fixing symbiosis. *Proceedings of the National Academy of Sciences*, *112*(49), 15238–15243. https://doi.org/10.1073/pnas.1500123112

Klimovich, A., & Bosch, T. C. G. (2024). Novel technologies uncover novel 'anti'-microbial peptides in Hydra shaping the species-specific microbiome. *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *379*(1901), 20230058. https://doi.org/10.1098/rstb.2023.0058

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: A multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), 559–574. https://doi.org/10.1002/prot.20921

Kryshtafovych, A., Antczak, M., Szachniuk, M., Zok, T., Kretsch, R. C., Rangan, R., Pham, P., Das, R., Robin, X., Studer, G., Durairaj, J., Eberhardt, J., Sweeney, A., Topf, M., Schwede, T., Fidelis, K., & Moult, J. (2023). New prediction categories in CASP15. *Proteins: Structure, Function, and Bioinformatics*, *91*(12), 1550–1557. https://doi.org/10.1002/prot.26515

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1011–1020. https://doi.org/10.1002/prot.25823

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1607–1617. https://doi.org/10.1002/prot.26237

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Lachat, J., Lextrait, G., Jouan, R., Boukherissa, A., Yokota, A., Jang, S., Ishigami, K., Futahashi, R., Cossard, R., Naquin, D., Costache, V., Augusto, L., Tissières, P., Biondi, E. G., Alunni, B., Timchenko, T., Ohbayashi, T., Kikuchi, Y., & Mergaert, P. (2024). Hundreds of antimicrobial peptides create a selective barrier for insect gut symbionts. *Proceedings of the National Academy of Sciences*, *121*(25), e2401802121. https://doi.org/10.1073/pnas.2401802121

Lambert, I., Paysant-Le Roux, C., Colella, S., & Martin-Magniette, M.-L. (2020). DiCoExpress: A tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. *Plant Methods*, *16*(1), 68. https://doi.org/10.1186/s13007-020-00611-7

Lamouche, F., Bonadé-Bottino, N., Mergaert, P., & Alunni, B. (2019). Symbiotic Efficiency of Spherical and Elongated Bacteroids in the Aeschynomene-Bradyrhizobium Symbiosis. *Frontiers in Plant Science*, *10*, 377. https://doi.org/10.3389/fpls.2019.00377

Lamouche, F., Gully, D., Chaumeret, A., Nouwen, N., Verly, C., Pierre, O., Sciallano, C., Fardoux, J., Jeudy, C., Szücs, A., Mondy, S., Salon, C., Nagy, I., Kereszt, A., Dessaux, Y., Giraud, E., Mergaert, P., & Alunni, B. (2019). Transcriptomic dissection of Bradyrhizobium sp. Strain ORS285 in symbiosis with Aeschynomene spp. Inducing different bacteroid morphotypes with contrasted symbiotic efficiency. *Environmental Microbiology*, *21*(9), 3244–3258. https://doi.org/10.1111/1462-2920.14292

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), Article 4. https://doi.org/10.1038/nmeth.1923

Laranjo, M., Alexandre, A., & Oliveira, S. (2014). Legume growth-promoting rhizobia: An overview on the *Mesorhizobium* genus. *Microbiological Research*, *169*(1), 2–17.

https://doi.org/10.1016/j.micres.2013.09.012

Lavin, M., Herendeen, P. S., & Wojciechowski, M. F. (2005). Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. *Systematic Biology*, *54*(4), 575–594. https://doi.org/10.1080/10635150590947131

Laviña, M., Pugsley, A. P., & Moreno, F. (1986). Identification, Mapping, Cloning and Characterization of a Gene (sbmA) Required for Microcin B17 Action on Escherichia coli K12. *Microbiology*, *132*(6), 1685–1693. https://doi.org/10.1099/00221287-132-6-1685

Lee, J., Jeong, B., Kim, J., Cho, J. H., Byeon, J. H., Lee, B. L., & Kim, J. K. (2024). Specialized digestive mechanism for an insect-bacterium gut symbiosis. *The ISME Journal*, *18*(1), wrad021. https://doi.org/10.1093/ismejo/wrad021

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

LeVier, K., Phillips, R. W., Grippe, V. K., Roop, R. M., II, & Walker, G. C. (2000). Similar Requirements of a Plant Symbiont and a Mammalian Pathogen for Prolonged Intracellular Survival. *Science*, *287*(5462), 2492–2493. https://doi.org/10.1126/science.287.5462.2492

LeVier, K., & Walker, G. C. (2001). Genetic Analysis of the Sinorhizobium meliloti BacA Protein: Differential Effects of Mutations on Phenotypes. *Journal of Bacteriology*, *183*(21), 6444–6453. https://doi.org/10.1128/jb.183.21.6444-6453.2001

Li, G., Laturnus, C., Ewers, C., & Wieler, L. H. (2005). Identification of Genes Required for Avian Escherichia coli Septicemia by Signature-Tagged Mutagenesis. *Infection and Immunity*, *73*(5), 2818–2827. https://doi.org/10.1128/iai.73.5.2818-2827.2005

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Lima, R. M., Rathod, B. B., Tiricz, H., Howan, D. H. O., Al Bouni, M. A., Jenei, S., Tímár, E., Endre, G., Tóth, G. K., & Kondorosi, É. (2022). Legume Plant Peptides as Sources of Novel Antimicrobial Molecules Against Human Pathogens. *Frontiers in Molecular Biosciences*, *9*. https://doi.org/10.3389/fmolb.2022.870460

Lind, P. A. (2019). Repeatability and Predictability in Experimental Evolution. In P. Pontarotti (Ed.), *Evolution, Origin of Life, Concepts and Methods* (pp. 57–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-30363-1_4

Liu, A., Ku, Y.-S., Contador, C. A., & Lam, H.-M. (2020). The Impacts of Domestication and Agricultural Practices on Legume Nutrient Acquisition Through Symbiosis With Rhizobia and Arbuscular Mycorrhizal Fungi. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.583954

Liu, J., Wu, T., Guo, Z., Hou, J., & Cheng, J. (2022). Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *Proteins: Structure, Function,*

and Bioinformatics, *90*(1), 58–72. https://doi.org/10.1002/prot.26186

Liu, Z., Chen, W., Jiao, S., Wang, X., Fan, M., Wang, E., & Wei, G. (2019). New Insight into the Evolution of Symbiotic Genes in Black Locust-Associated Rhizobia. *Genome Biology and Evolution*, *11*(7), 1736–1750. https://doi.org/10.1093/gbe/evz116

Lohse, K., Gutierrez, A., & Kaltz, O. (2006). EXPERIMENTAL EVOLUTION OF RESISTANCE IN PARAMECIUM CAUDATUM AGAINST THE BACTERIAL PARASITE HOLOSPORA UNDULATA. *Evolution*, *60*(6), 1177–1186. https://doi.org/10.1111/j.0014-3820.2006.tb01196.x

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lupas, A. N., Pereira, J., Alva, V., Merino, F., Coles, M., & Hartmann, M. D. (2021). The breakthrough in protein structure prediction. *Biochemical Journal*, *478*(10), 1885–1890. https://doi.org/10.1042/BCJ20200963

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., & Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, *95*(6), 3140–3145. https://doi.org/10.1073/pnas.95.6.3140

Mallet, J., Besansky, N., & Hahn, M. W. (2016). How reticulated are species? *BioEssays*, *38*(2), 140–149. https://doi.org/10.1002/bies.201500149

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Marlow, V. L., Haag, A. F., Kobayashi, H., Fletcher, V., Scocchi, M., Walker, G. C., & Ferguson, G. P. (2009). Essential role for the BacA protein in the uptake of a truncated eukaryotic peptide in Sinorhizobium meliloti. *Journal of Bacteriology*, *191*(5), 1519–1527. https://doi.org/10.1128/JB.01661-08

Maróti, G., Downie, J. A., & Kondorosi, É. (2015). Plant cysteine-rich peptides that inhibit pathogen growth and control rhizobial differentiation in legume nodules. *Current Opinion in Plant Biology*, *26*, 57–63. https://doi.org/10.1016/j.pbi.2015.05.031

Maróti, G., Kereszt, A., Kondorosi, É., & Mergaert, P. (2011). Natural roles of antimicrobial peptides in microbes, plants and animals. *Research in Microbiology*, *162*(4), 363–374. https://doi.org/10.1016/j.resmic.2011.02.005

Maróti, G., & Kondorosi, É. (2014). Nitrogen-fixing Rhizobium-legume symbiosis: Are polyploidy and host peptide-governed symbiont differentiation general principles of endosymbiosis? *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00326

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing

reads. *EMBnet.Journal*, *17*(1), Article 1. https://doi.org/10.14806/ej.17.1.200

Martin, W. F., Garg, S., & Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140330. https://doi.org/10.1098/rstb.2014.0330

Martínez-Romero, E. (2009). Coevolution in Rhizobium-Legume Symbiosis? *DNA and Cell Biology*, *28*(8), 361–370. https://doi.org/10.1089/dna.2009.0863

Maruya, J., & Saeki, K. (2010). The bacA gene homolog, mlr7400, in Mesorhizobium loti MAFF303099 is dispensable for symbiosis with Lotus japonicus but partially capable of supporting the symbiotic function of bacA in Sinorhizobium meliloti. *Plant & Cell Physiology*, *51*(9), 1443–1452. https://doi.org/10.1093/pcp/pcq114

Mattiuzzo, M., Bandiera, A., Gennaro, R., Benincasa, M., Pacor, S., Antcheva, N., & Scocchi, M. (2007). Role of the Escherichia coli SbmA in the antimicrobial activity of proline-rich peptides. *Molecular Microbiology*, *66*(1), 151–163. https://doi.org/10.1111/j.1365-2958.2007.05903.x

Mehnaz, S. (2011). Plant Growth-Promoting Bacteria Associated with Sugarcane. In D. K. Maheshwari (Ed.), *Bacteria in Agrobiology: Crop Ecosystems* (pp. 165–187). Springer. https://doi.org/10.1007/978-3-642-18357-7_7

Menegat, S., Ledo, A., & Tirado, R. (2022). Greenhouse gas emissions from global production and use of nitrogen synthetic fertilisers in agriculture. *Scientific Reports*, *12*(1), 14490. https://doi.org/10.1038/s41598-022-18773-w

Mergaert, P. (2018). Role of antimicrobial peptides in controlling symbiotic bacterial populations. *Natural Product Reports*, *35*(4), 336–356. https://doi.org/10.1039/C7NP00056A

Mergaert, P., Nikovics, K., Kelemen, Z., Maunoury, N., Vaubert, D., Kondorosi, A., & Kondorosi, E. (2003). A Novel Family in Medicago truncatula Consisting of More Than 300 Nodule-Specific Genes Coding for Small, Secreted Polypeptides with Conserved Cysteine Motifs,. *Plant Physiology*, *132*(1), 161–173. https://doi.org/10.1104/pp.102.018192

Mergaert, P., Uchiumi, T., Alunni, B., Evanno, G., Cheron, A., Catrice, O., Mausset, A.-E., Barloy-Hubler, F., Galibert, F., Kondorosi, A., & Kondorosi, E. (2006). Eukaryotic control on bacterial cell cycle and differentiation in the Rhizobium–legume symbiosis. *Proceedings of the National Academy of Sciences*, *103*(13), 5230–5235. https://doi.org/10.1073/pnas.0600912103

Mergaert, P., Van Montagu, M., & Holsters, M. (1997). Molecular mechanisms of Nod factor diversity. *Molecular Microbiology*, *25*(5), 811–817. https://doi.org/10.1111/j.1365-2958.1997.mmi526.x

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mishra, N. C., & Tatum, E. L. (1973). Non-Mendelian Inheritance of DNA-Induced Inositol

Independence in Neurospora. *Proceedings of the National Academy of Sciences*, *70*(12), 3875–3879. https://doi.org/10.1073/pnas.70.12.3875

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121. https://doi.org/10.1093/nar/gkt263

Modi, V., & Dunbrack Jr., R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, *84*(S1), 260–281. https://doi.org/10.1002/prot.25048

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2023). *Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses* (p. 2023.09.19.558401). bioRxiv. https://doi.org/10.1101/2023.09.19.558401

Montiel, J., Downie, J. A., Farkas, A., Bihari, P., Herczeg, R., Bálint, B., Mergaert, P., Kereszt, A., & Kondorosi, É. (2017). Morphotype of bacteroids in different legumes correlates with the number and type of symbiotic NCR peptides. *Proceedings of the National Academy of Sciences*, *114*(19), 5041–5046. https://doi.org/10.1073/pnas.1704217114

Montiel, J., Szűcs, A., Boboescu, I. Z., Gherman, V. D., Kondorosi, É., & Kereszt, A. (2016). Terminal Bacteroid Differentiation Is Associated With Variable Morphological Changes in Legume Species Belonging to the Inverted Repeat-Lacking Clade. *Molecular Plant-Microbe Interactions®*, *29*(3), 210–219. https://doi.org/10.1094/MPMI-09-15-0213-R

Monzon, V., Paysan-Lafosse, T., Wood, V., & Bateman, A. (2022). Reciprocal best structure hits: Using AlphaFold models to discover distant homologues. *Bioinformatics Advances*, *2*(1), vbac072. https://doi.org/10.1093/bioadv/vbac072

Morgado, S., Ramos, N. de V., Pereira, B. B. do N., Freitas, F., da Fonseca, É. L., & Vicente, A. C. (2022). Multidrug-resistant Mycolicibacterium fortuitum infection in a companion cat (Felis silvestris catus) in Brazil. *Access Microbiology*, *4*(2), 000317. https://doi.org/10.1099/acmi.0.000317

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics*, *84*(S1), 4–14. https://doi.org/10.1002/prot.25064

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics*, *86*(S1), 7–15. https://doi.org/10.1002/prot.25415

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, *23*(3), ii–iv. https://doi.org/10.1002/prot.340230303

Mutalipassi, M., Riccio, G., Mazzella, V., Galasso, C., Somma, E., Chiarore, A., de Pascale, D., & Zupo, V. (2021). Symbioses of Cyanobacteria in Marine Environments: Ecological Insights and Biotechnological Perspectives. *Marine Drugs*, *19*(4), Article 4. https://doi.org/10.3390/md19040227

Nagy, L. G., Ohm, R. A., Kovács, G. M., Floudas, D., Riley, R., Gácser, A., Sipiczki, M., Davis, J. M., Doty, S. L., de Hoog, G. S., Lang, B. F., Spatafora, J. W., Martin, F. M., Grigoriev, I. V., & Hibbett, D. S. (2014). Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nature Communications*, *5*(1), 4471. https://doi.org/10.1038/ncomms5471

Nair, R. R., Vasse, M., Wielgoss, S., Sun, L., Yu, Y.-T. N., & Velicer, G. J. (2019). Bacterial predator-prey coevolution accelerates genome evolution and selects on virulence-associated prey defences. *Nature Communications*, *10*(1), 4301. https://doi.org/10.1038/s41467-019-12140-6

Nallu, S., Silverstein, K. A. T., Zhou, P., Young, N. D., & VandenBosch, K. A. (2014). Patterns of divergence of a large family of nodule cysteine-rich peptides in accessions of Medicago truncatula. *The Plant Journal*, *78*(4), 697–705. https://doi.org/10.1111/tpj.12506

Ndagi, U., Falaki, A. A., Abdullahi, M., Lawal, M. M., & Soliman, M. E. (2020). Antibiotic resistance: Bioinformatics-based understanding as a functional strategy for drug design. *RSC Advances*, *10*(31), 18451–18468. https://doi.org/10.1039/D0RA01484B

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nesse, R. M., Stearns, S. C., & Omenn, G. S. (2006). Medicine Needs Evolution. *Science*, *311*(5764), 1071–1071. https://doi.org/10.1126/science.1125956

Newcomb, W., & Wood, S. M. (1986). Fine structure of nitrogen-fixing leguminous root nodules from the Canadian Arctic. *Nordic Journal of Botany*, *6*(5), 609–626. https://doi.org/10.1111/j.1756-1051.1986.tb00461.x

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nicoud, Q., Barrière, Q., Busset, N., Dendene, S., Travin, D., Bourge, M., Le Bars, R., Boulogne, C., Lecroël, M., Jenei, S., Kereszt, A., Kondorosi, E., Biondi, E. G., Timchenko, T., Alunni, B., & Mergaert, P. (2021). Sinorhizobium meliloti Functions Required for Resistance to Antimicrobial NCR Peptides and Bacteroid Differentiation. *mBio*, *12*(4), e00895-21. https://doi.org/10.1128/mBio.00895-21

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Oberg, N., Zallot, R., & Gerlt, J. A. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *Journal of Molecular Biology*, *435*(14), 168018. https://doi.org/10.1016/j.jmb.2023.168018

Oldroyd, G. E. D. (2013). Speak, friend, and enter: Signalling systems that promote beneficial symbiotic associations in plants. *Nature Reviews Microbiology*, *11*(4), Article 4. https://doi.org/10.1038/nrmicro2990

Oono, R., & Denison, R. F. (2010). Comparing Symbiotic Efficiency between Swollen versus Nonswollen Rhizobial Bacteroids. *Plant Physiology*, *154*(3), 1541–1548. https://doi.org/10.1104/pp.110.163436

Oono, R., Schmitt, I., Sprent, J. I., & Denison, R. F. (2010). Multiple evolutionary origins of legume traits leading to extreme rhizobial differentiation. *New Phytologist*, *187*(2), 508–520. https://doi.org/10.1111/j.1469-8137.2010.03261.x

Ourèye Sy, M., Hocher, V., Gherbi, H., Laplaze, L., Auguy, F., Bogusz, D., & Franche, C. (2007). The cell-cycle promoter cdc2aAt from Arabidopsis thaliana is induced in the lateral roots of the actinorhizal tree Allocasuarina verticillata during the early stages of the symbiotic interaction with Frankia. *Physiologia Plantarum*, *130*(3), 409–417. https://doi.org/10.1111/j.1399-3054.2007.00884.x

Pan, H., & Wang, D. (2017). Nodule cysteine-rich peptides maintain a working balance during nitrogen-fixing symbiosis. *Nature Plants*, *3*(5), Article 5. https://doi.org/10.1038/nplants.2017.48

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), Article 4. https://doi.org/10.1038/nmeth.4197

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1.1-3.1.8. https://doi.org/10.1002/0471250953.bi0301s42

Penterman, J., Abo, R. P., De Nisco, N. J., Arnold, M. F. F., Longhi, R., Zanda, M., & Walker, G. C. (2014). Host plant peptides elicit a transcriptional response to control the Sinorhizobium meliloti cell cycle during symbiosis. *Proceedings of the National Academy of Sciences*, *111*(9), 3561–3566. https://doi.org/10.1073/pnas.1400450111

Peoples, M. B., Brockwell, J., Herridge, D. F., Rochester, I. J., Alves, B. J. R., Urquiaga, S., Boddey, R. M., Dakora, F. D., Bhattarai, S., Maskey, S. L., Sampet, C., Rerkasem, B., Khan, D. F., Hauggaard-Nielsen, H., & Jensen, E. S. (2009). The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. *Symbiosis*, *48*(1), 1–17. https://doi.org/10.1007/BF03179980

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), Article 10. https://doi.org/10.1038/nmeth.1701

Pini, F., Nisco, N. J. D., Ferri, L., Penterman, J., Fioravanti, A., Brilli, M., Mengoni, A., Bazzicalupo, M., Viollier, P. H., Walker, G. C., & Biondi, E. G. (2015). Cell Cycle Control by the Master Regulator CtrA in Sinorhizobium meliloti. *PLOS Genetics*, *11*(5), e1005232. https://doi.org/10.1371/journal.pgen.1005232

Poole, P., Ramachandran, V., & Terpolilli, J. (2018). Rhizobia: From saprophytes to endosymbionts. *Nature Reviews Microbiology*, *16*(5), 291–303. https://doi.org/10.1038/nrmicro.2017.171

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*, *38*(10), 2943–2945. https://doi.org/10.1093/bioinformatics/btac166

Rahimlou, S., Bahram, M., & Tedersoo, L. (2021). Phylogenomics reveals the evolution of root nodulating alpha- and beta-Proteobacteria (rhizobia). *Microbiological Research*, *250*, 126788. https://doi.org/10.1016/j.micres.2021.126788

Raina, J.-B., Eme, L., Pollock, F. J., Spang, A., Archibald, J. M., & Williams, T. A. (2018). Symbiosis in the microbial world: From ecology to genome evolution. *Biology Open*, *7*(2), bio032524. https://doi.org/10.1242/bio.032524

Raul, B., Bhattacharjee, O., Ghosh, A., Upadhyay, P., Tembhare, K., Singh, A., Shaheen, T., Ghosh, A. K., Torres-Jerez, I., Krom, N., Clevenger, J., Udvardi, M., E. Scheffler, B., Ozias Akins, P., Dutta Sharma, R., Bandyopadhyay, K., Gaur, V., Kumar, S., & Sinharoy, S. (2021). Microscopic and transcriptomic analyses of Dalbergoid legume peanut reveal a divergent evolution leading to Nod Factor dependent epidermal crack-entry and terminal bacteroid differentiation. *Molecular Plant-Microbe Interactions*®. https://doi.org/10.1094/MPMI-05-21-0122-R

Reidenberg, J. S. (2007). Anatomical adaptations of aquatic mammals. *The Anatomical Record*, *290*(6), 507–513. https://doi.org/10.1002/ar.20541

Ren, G. (2018). *The evolution of determinate and indeterminate nodules within the Papilionoideae subfamily*. https://doi.org/10.18174/429101

Ritchie, M. E., & Tilman, D. (1995). Responses of Legumes to Herbivores and Nutrients During Succession on a Nitrogen-Poor Soil. *Ecology*, *76*(8), 2648–2655. https://doi.org/10.2307/2265835

Rivas, R., Velázquez, E., Willems, A., Vizcaíno, N., Subba-Rao, N. S., Mateos, P. F., Gillis, M., Dazzo, F. B., & Martínez-Molina, E. (2002). A New Species of Devosia That Forms a Unique Nitrogen-Fixing Root-Nodule Symbiosis with the Aquatic Legume Neptunia natans (L.f.) Druce. *Applied and Environmental Microbiology*, *68*(11), 5217–5222. https://doi.org/10.1128/AEM.68.11.5217-5222.2002

Rivas, R., Willems, A., Subba-Rao, N. S., Mateos, P. F., Dazzo, F. B., Kroppenstedt, R. M., Martínez-Molina, E., Gillis, M., & Velázquez, E. (2003). Description of *Devosia neptuniae* sp. Nov. That Nodulates and Fixes Nitrogen in Symbiosis with *Neptunia natans*, an Aquatic Legume from India. *Systematic and Applied Microbiology*, *26*(1), 47–53. https://doi.org/10.1078/072320203322337308

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rosca, V., Duca, M., de Groot, M. T., & Koper, M. T. M. (2009). Nitrogen Cycle Electrocatalysis. *Chemical Reviews*, *109*(6), 2209–2244. https://doi.org/10.1021/cr8003696

Roux, B., Rodde, N., Jardinaud, M.-F., Timmers, T., Sauviac, L., Cottret, L., Carrère, S., Sallet, E., Courcelle, E., Moreau, S., Debellé, F., Capela, D., de Carvalho-Niebel, F., Gouzy, J., Bruand, C., & Gamas, P. (2014). An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing. *The Plant Journal*, *77*(6), 817–837. https://doi.org/10.1111/tpj.12442

Roy, P., Achom, M., Wilkinson, H., Lagunas, B., & Gifford, M. L. (2020). Symbiotic Outcome Modified by the Diversification from 7 to over 700 Nodule-Specific Cysteine-Rich Peptides. *Genes*, *11*(4), Article 4. https://doi.org/10.3390/genes11040348

Runti, G., Lopez Ruiz, M. del C., Stoilova, T., Hussain, R., Jennions, M., Choudhury, H. G., Benincasa, M., Gennaro, R., Beis, K., & Scocchi, M. (2013). Functional Characterization of SbmA, a Bacterial Inner Membrane Transporter Required for Importing the Antimicrobial Peptide Bac7(1-35). *Journal of Bacteriology*, *195*(23), 5343–5351. https://doi.org/10.1128/jb.00818-13

Sackton, T. B., Grayson, P., Cloutier, A., Hu, Z., Liu, J. S., Wheeler, N. E., Gardner, P. P., Clarke, J. A., Baker, A. J., Clamp, M., & Edwards, S. V. (2019). Convergent regulatory evolution and loss of flight in paleognathous birds. *Science*, *364*(6435), 74–78. https://doi.org/10.1126/science.aat7244

Salgado, M. G., Demina, I. V., Maity, P. J., Nagchowdhury, A., Caputo, A., Krol, E., Loderer, C., Muth, G., Becker, A., & Pawlowski, K. (2022). Legume NCRs and nodule-specific defensins of actinorhizal plants—Do they share a common origin? *PLOS ONE*, *17*(8), e0268683. https://doi.org/10.1371/journal.pone.0268683

Sandler, I., Medalia, O., & Aharoni, A. (2013). Experimental analysis of co-evolution within protein complexes: The yeast exosome as a model. *Proteins: Structure, Function, and Bioinformatics*, *81*(11), 1997–2006. https://doi.org/10.1002/prot.24360

Sankari, S., Babu, V. M. P., Bian, K., Alhhazmi, A., Andorfer, M. C., Avalos, D. M., Smith, T. A., Yoon, K., Drennan, C. L., Yaffe, M. B., Lourido, S., & Walker, G. C. (2022). A haem-sequestering plant peptide promotes iron uptake in symbiotic bacteria. *Nature Microbiology*, *7*(9), Article 9. https://doi.org/10.1038/s41564-022-01192-y

Sato, S., Kaneko, T., Nakamura, Y., Asamizu, E., Kato, T., & Tabata, S. (2008). Structural and Comparative Genome Analysis of Lotus japonicus. In F. D. Dakora, S. B. M. Chimphango, A. J. Valentine, C. Elmerich, & W. E. Newton (Eds.), *Biological Nitrogen Fixation: Towards Poverty Alleviation through Sustainable Agriculture* (pp. 217–219). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8252-8_80

Schlieter, H., Stark, J., Burwitz, M., & Braun, R. (2019). Terminology for Evolving Design Artifacts. *Wirtschaftsinformatik 2019 Proceedings*. https://aisel.aisnet.org/wi2019/track03/papers/8

Schrire, B. D., Lavin, M., Barker, N. P., & Forest, F. (2009). Phylogeny of the tribe Indigofereae (Leguminosae–Papilionoideae): Geographically structured more in succulent-rich and temperate settings than in grass-rich environments. *American Journal of Botany*, *96*(4), 816–852. https://doi.org/10.3732/ajb.0800185

Sen, A., Beauchemin, N., Bruce, D., Chain, P., Chen, A., Walston Davenport, K., Deshpande, S., Detter, C., Furnholm, T., Ghodbhane-Gtari, F., Goodwin, L., Gtari, M., Han, C., Han, J., Huntemann, M., Ivanova, N., Kyrpides, N., Land, M. L., Markowitz, V., … Tisa, L. S. (2013). Draft Genome Sequence of Frankia sp. Strain QA3, a Nitrogen-Fixing Actinobacterium Isolated from the Root Nodule of Alnus nitida. *Genome Announcements*, *1*(2), 10.1128/genomea.00103-13. https://doi.org/10.1128/genomea.00103-13

Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C.-C., & Chain, P. S. G. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific Reports*, *10*(1), 1723. https://doi.org/10.1038/s41598-020-58356-1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Silva, F. G., Filho, R. M. M., Martins, L. S. S., Ramos, R. da S., & Silva, G. C. (2022). *Plastid marker-based phylogeny reveals insights into relationships among Papilionoideae species*. https://doi.org/10.21203/rs.3.rs-2347656/v1

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Singh, R. J., Chung, G. H., & Nelson, R. L. (2007). Landmark research in legumes. *Genome*, *50*(6), 525–537. https://doi.org/10.1139/G07-037

Skorupska, A., Janczarek, M., Marczak, M., Mazur, A., & Król, J. (2006). Rhizobial exopolysaccharides: Genetic control and symbiotic functions. *Microbial Cell Factories*, *5*(1), 7. https://doi.org/10.1186/1475-2859-5-7

Slotboom, D. J., Ettema, T. W., Nijland, M., & Thangaratnarajah, C. (2020). Bacterial multi-solute transporters. *FEBS Letters*, *594*(23), 3898–3907. https://doi.org/10.1002/1873-3468.13912

Smil, V. (1999). Detonator of the population explosion. *Nature*, *400*(6743), 415–415. https://doi.org/10.1038/22672

Smith, N. T., Boukherissa, A., Antaya, K., Howe, G. W., Vega, R. C. R. de la, Shykoff, J. A., Alunni, B., & diCenzo, G. C. (2024). *Taxonomic distribution of SbmA/BacA and BacA-like antimicrobial peptide transporters suggests independent recruitment and convergent evolution in host-microbe interactions* (p. 2024.02.25.581009). bioRxiv. https://doi.org/10.1101/2024.02.25.581009

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Sniegowski, P. D., & Lenski, R. E. (1995). MUTATION AND ADAPTATION: The Directed Mutation Controversy in Evolutionary Perspective. *Annual Review of Ecology, Evolution, and Systematics*, *26*(Volume 26,), 553–578. https://doi.org/10.1146/annurev.es.26.110195.003005

Song, L., & Florea, L. (2015). Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*(1), s13742-015-0089-y. https://doi.org/10.1186/s13742-015-0089-y

Sonnhammer, E. L. L., Eddy, S. R., & Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, *28*(3), 405–420. https://doi.org/10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stępkowski, T., Banasiewicz, J., Granada, C. E., Andrews, M., & Passaglia, L. M. P. (2018). Phylogeny and Phylogeography of Rhizobial Symbionts Nodulating Legumes of the Tribe Genisteae. *Genes*, *9*(3), 163. https://doi.org/10.3390/genes9030163

Szerencsés, B., Gácser, A., Endre, G., Domonkos, I., Tiricz, H., Vágvölgyi, C., Szolomajer, J., Howan, D. H. O., Tóth, G. K., Pfeiffer, I., & Kondorosi, É. (2021). Symbiotic NCR Peptide Fragments Affect the Viability, Morphology and Biofilm Formation of Candida Species. *International Journal of Molecular Sciences*, *22*(7), Article 7. https://doi.org/10.3390/ijms22073666

Tan, X.-J., Cheng, Y., Li, Y.-X., Li, Y.-G., & Zhou, J.-C. (2009). BacA is indispensable for successful Mesorhizobium–Astragalus symbiosis. *Applied Microbiology and Biotechnology*, *84*(3), 519–526. https://doi.org/10.1007/s00253-009-1959-y

Tatusova, T. A., & Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, *174*(2), 247–250. https://doi.org/10.1111/j.1574-6968.1999.tb13575.x

Thomas, G. W. C., & Hahn, M. W. (2015). Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. *Molecular Biology and Evolution*, *32*(5), 1232–1236. https://doi.org/10.1093/molbev/msv013

Thompson, A. W., Foster, R. A., Krupke, A., Carter, B. J., Musat, N., Vaulot, D., Kuypers, M. M. M., & Zehr, J. P. (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science*, *337*(6101), 1546–1550. https://doi.org/10.1126/science.1222700

Tiricz, H., Szűcs, A., Farkas, A., Pap, B., Lima, R. M., Maróti, G., Kondorosi, É., & Kereszt, A. (2013). Antimicrobial Nodule-Specific Cysteine-Rich Peptides Induce Membrane Depolarization-Associated Changes in the Transcriptome of Sinorhizobium meliloti. *Applied and Environmental Microbiology*, *79*(21), 6737–6746. https://doi.org/10.1128/AEM.01791-13

Travin, D. Y., Jouan, R., Vigouroux, A., Inaba-Inoue, S., Lachat, J., Haq, F., Timchenko, T., Sutormin, D., Dubiley, S., Beis, K., Moréra, S., Severinov, K., & Mergaert, P. (2023). Dual-Uptake Mode of the Antibiotic Phazolicin Prevents Resistance Acquisition by Gram-Negative Bacteria. *mBio*, *14*(2), e0021723. https://doi.org/10.1128/mbio.00217-23

Travin, D. Y., Vigouroux, A., Inaba-Inoue, S., Qu, F., Jouan, R., Lachat, J., Sutormin, D., Dubiley, S., Beis, K., Moréra, S., Severinov, K., & Mergaert, P. (2022). *The antibiotic phazolicin displays a dual mode of uptake in Gram-negative bacteria* (p. 2022.04.27.489825). bioRxiv. https://doi.org/10.1101/2022.04.27.489825

Twomey, E., Melo-Sampaio, P., Schulte, L. M., Bossuyt, F., Brown, J. L., & Castroviejo-Fisher, S. (2023). Multiple Routes to Color Convergence in a Radiation of Neotropical Poison Frogs. *Systematic Biology*, *72*(6), 1247–1261. https://doi.org/10.1093/sysbio/syad051

Ujvari, B., Casewell, N. R., Sunagar, K., Arbuckle, K., Wüster, W., Lo, N., O'Meally, D., Beckmann, C., King, G. F., Deplazes, E., & Madsen, T. (2015). Widespread convergence in toxin resistance by predictable molecular evolution. *Proceedings of the National Academy of Sciences*, *112*(38), 11911–11916. https://doi.org/10.1073/pnas.1511706112

Val-Calvo, J., & Vázquez-Boland, J. A. (2023). Mycobacteriales taxonomy using network analysis-aided, context-uniform phylogenomic approach for non-subjective genus demarcation. *mBio*, *14*(5), e02207-23. https://doi.org/10.1128/mbio.02207-23

Van de Velde, W., Zehirov, G., Szatmari, A., Debreczeny, M., Ishihara, H., Kevei, Z., Farkas, A., Mikulass, K., Nagy, A., Tiricz, H., Satiat-Jeunemaître, B., Alunni, B., Bourge, M., Kucho, K., Abe, M., Kereszt, A., Maroti, G., Uchiumi, T., Kondorosi, E., & Mergaert, P. (2010). Plant Peptides Govern Terminal Differentiation of Bacteria in Symbiosis. *Science*, *327*(5969), 1122–1126. https://doi.org/10.1126/science.1184057

Van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, *30*(1), 121–141. https://doi.org/10.1137/040608635

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, *42*(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Varshney, R. K., & Kudapa, H. (2013). Legume biology: The basis for crop improvement. *Functional Plant Biology*, *40*(12), v–viii. https://doi.org/10.1071/FPv40n12_FO

Vasse, J., de Billy, F., Camut, S., & Truchet, G. (1990). Correlation between ultrastructural differentiation of bacteroids and nitrogen fixation in alfalfa nodules. *Journal of Bacteriology*, *172*(8), 4295–4306. https://doi.org/10.1128/jb.172.8.4295-4306.1990

Vincent, J. M. (1970). A manual for the practical study of the root-nodule bacteria. *A Manual for the Practical Study of the Root-Nodule Bacteria.* https://www.cabdirect.org/cabdirect/abstract/19710700726

Voisin, A.-S., Guéguen, J., Huyghe, C., Jeuffroy, M.-H., Magrini, M.-B., Meynard, J.-M., Mougel, C., Pellerin, S., & Pelzer, E. (2014). Legumes for feed, food, biomaterials and bioenergy in Europe: A review. *Agronomy for Sustainable Development*, *34*(2), 361–380. https://doi.org/10.1007/s13593-013-0189-y

Vondenhoff, G. H. M., Blanchaert, B., Geboers, S., Kazakov, T., Datsenko, K. A., Wanner, B. L., Rozenski, J., Severinov, K., & Van Aerschot, A. (2011). Characterization of Peptide Chain Length and Constituency Requirements for YejABEF-Mediated Uptake of Microcin C Analogues. *Journal of Bacteriology*, *193*(14), 3618–3623. https://doi.org/10.1128/jb.00172-11

Voronina, O. L., Kunda, M. S., Ryzhova, N. N., Aksenova, E. I., Semenov, A. N., Lasareva, A. V., Amelina, E. L., Chuchalin, A. G., Lunin, V. G., & Gintsburg, A. L. (2015). The Variability of the Order Burkholderiales Representatives in the Healthcare Units. *BioMed Research International*, *2015*, e680210. https://doi.org/10.1155/2015/680210

Walton, J. H., Kontra-Kováts, G., Green, R. T., Domonkos, Á., Horváth, B., Brear, E. M., Franceschetti, M., Kaló, P., & Balk, J. (2020). The Medicago truncatula Vacuolar iron Transporter-Like proteins VTL4 and VTL8 deliver iron to symbiotic bacteria at different stages of the infection process. *New Phytologist*, *228*(2), 651–666. https://doi.org/10.1111/nph.16735

Wang, E. T. (2019). Symbiosis Between Rhizobia and Legumes. In E. T. Wang, C. F. Tian, W. F. Chen, J. P. W. Young, & W. X. Chen (Eds.), *Ecology and Evolution of Rhizobia: Principles and Applications* (pp. 3–19). Springer. https://doi.org/10.1007/978-981-32-9555-1_1

Wang, L., Rubio, M. C., Xin, X., Zhang, B., Fan, Q., Wang, Q., Ning, G., Becana, M., & Duanmu, D. (2019). CRISPR/Cas9 knockout of leghemoglobin genes in Lotus japonicus uncovers their synergistic roles in symbiotic nitrogen fixation. *New Phytologist*, *224*(2), 818–832. https://doi.org/10.1111/nph.16077

Wang, L., Xia, M., Wang, H., Huang, K., Qian, C., Maravelias, C. T., & Ozin, G. A. (2018). Greening Ammonia toward the Solar Ammonia Refinery. *Joule*, *2*(6), 1055–1074. https://doi.org/10.1016/j.joule.2018.04.017

Wang, Z., Bie, P., Cheng, J., Lu, L., Cui, B., & Wu, Q. (2016). The ABC transporter YejABEF is required for resistance to antimicrobial peptides and the virulence of Brucella melitensis. *Scientific Reports*, *6*(1), 31876. https://doi.org/10.1038/srep31876

Waughman, G. J., & Bellamy, D. J. (1980). Nitrogen Fixation and the Nitrogen Balance in Peatland Ecosystems. *Ecology*, *61*(5), 1185–1198. https://doi.org/10.2307/1936837

Wehmeier, S., Arnold, M. F. F., Marlow, V. L., Aouida, M., Myka, K. K., Fletcher, V., Benincasa, M., Scocchi, M., Ramotar, D., & Ferguson, G. P. (2010). Internalization of a thiazole-modified peptide in Sinorhizobium meliloti occurs by BacA-dependent and -independent mechanisms. *Microbiology*, *156*(9), 2702–2713. https://doi.org/10.1099/mic.0.039909-0

Wei, F., Liu, Y., Zhou, D., Zhao, W., Chen, Z., Chen, D., Li, Y., & Zhang, X.-X. (2022). Transcriptomic Identification of a Unique Set of Nodule-Specific Cysteine-Rich Peptides Expressed in the Nitrogen-Fixing Root Nodule of Astragalus sinicus. *Molecular Plant-Microbe Interactions®*, *35*(10), 893–905. https://doi.org/10.1094/MPMI-03-22-0054-R

WELDON, W. F. R. (1902). MENDEL'S LAWS OF ALTERNATIVE INHERITANCE IN PEAS. *Biometrika*, *1*(2), 228–233. https://doi.org/10.1093/biomet/1.2.228

Wells, D. H., & Long, S. R. (2002). The Sinorhizobium meliloti stringent response affects multiple aspects of symbiosis. *Molecular Microbiology*, *43*(5), 1115–1127. https://doi.org/10.1046/j.1365-2958.2002.02826.x

Wheat, C. W., Vogel, H., Wittstock, U., Braby, M. F., Underwood, D., & Mitchell-Olds, T. (2007). The genetic basis of a plant–insect coevolutionary key innovation. *Proceedings of the National Academy of Sciences*, *104*(51), 20427–20431. https://doi.org/10.1073/pnas.0706229104

Willems, A. (2006). The taxonomy of rhizobia: An overview. *Plant and Soil*, *287*(1), 3–14. https://doi.org/10.1007/s11104-006-9058-7

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics (Oxford, England)*, *28*(7), 1033–1034. https://doi.org/10.1093/bioinformatics/bts079

Xiao, T. T., Schilderink, S., Moling, S., Deinum, E. E., Kondorosi, E., Franssen, H., Kulikova, O., Niebel, A., & Bisseling, T. (2014). Fate map of Medicago truncatula root nodules. *Development*, *141*(18), 3517–3528. https://doi.org/10.1242/dev.110775

Xu, D., Yang, Y., Gong, D., Chen, X., Jin, K., Jiang, H., Yu, W., Li, J., Zhang, J., & Pan, W. (2023). GFAP: Ultrafast and accurate gene functional annotation software for plants. *Plant Physiology*, *193*(3), 1745–1748. https://doi.org/10.1093/plphys/kiad393

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, *26*(7), 889–895. https://doi.org/10.1093/bioinformatics/btq066

Xu, Q., & Dunbrack, R. L., Jr. (2012). Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics*, *28*(21), 2763–2772. https://doi.org/10.1093/bioinformatics/bts533

Xu, S., Wang, J., Guo, Z., He, Z., & Shi, S. (2020). Genomic Convergence in the Adaptation to Extreme Environments. *Plant Communications*, *1*(6). https://doi.org/10.1016/j.xplc.2020.100117

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., Mayer, K. F. X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., … Roe, B. A. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, *480*(7378), Article 7378. https://doi.org/10.1038/nature10625

Yu, T., & Zhuang, Q. (2020). Modeling biological nitrogen fixation in global natural terrestrial ecosystems. *Biogeosciences*, *17*(13), 3643–3657. https://doi.org/10.5194/bg-17-3643-2020

Zakon, H. H. (2002). Convergent Evolution on the Molecular Level. *Brain Behavior and Evolution*, *59*(5–6), 250–261. https://doi.org/10.1159/000063562

Zallot, R., Oberg, N., & Gerlt, J. A. (2019). The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*, *58*(41), 4169–4182. https://doi.org/10.1021/acs.biochem.9b00735

Zehr, J. P., Waterbury, J. B., Turner, P. J., Montoya, J. P., Omoregie, E., Steward, G. F., Hansen, A., & Karl, D. M. (2001). Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean. *Nature*, *412*(6847), 635–638. https://doi.org/10.1038/35088063

Zhang, C., Shine, M., Pyle, A. M., & Zhang, Y. (2022). US-align: Universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, *19*(9), 1109–1115. https://doi.org/10.1038/s41592-022-01585-1

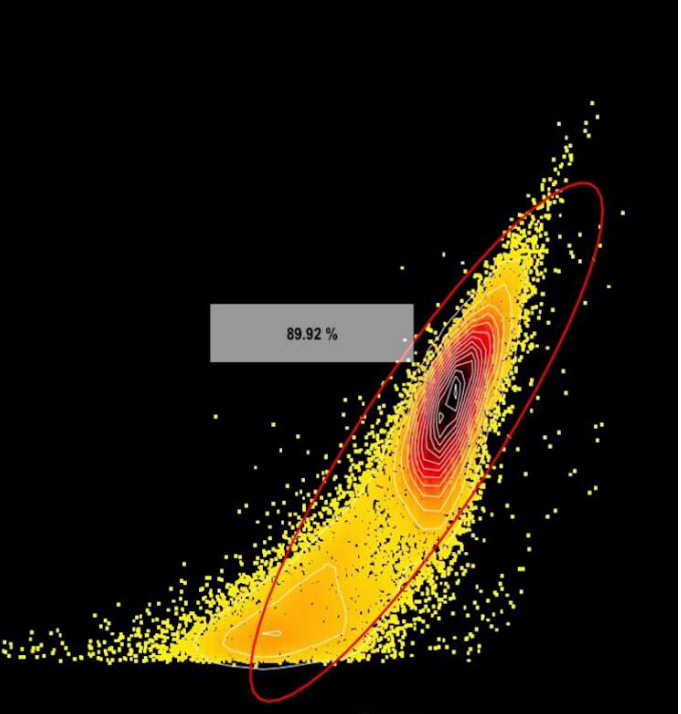
Zhang, S., Wang, T., Lima, R. M., Pettkó-Szandtner, A., Kereszt, A., Downie, J. A., & Kondorosi, E. (2023). Widely conserved AHL transcription factors are essential for NCR gene expression and nodule development in Medicago. *Nature Plants*, *9*(2), Article 2. https://doi.org/10.1038/s41477-022-01326-4

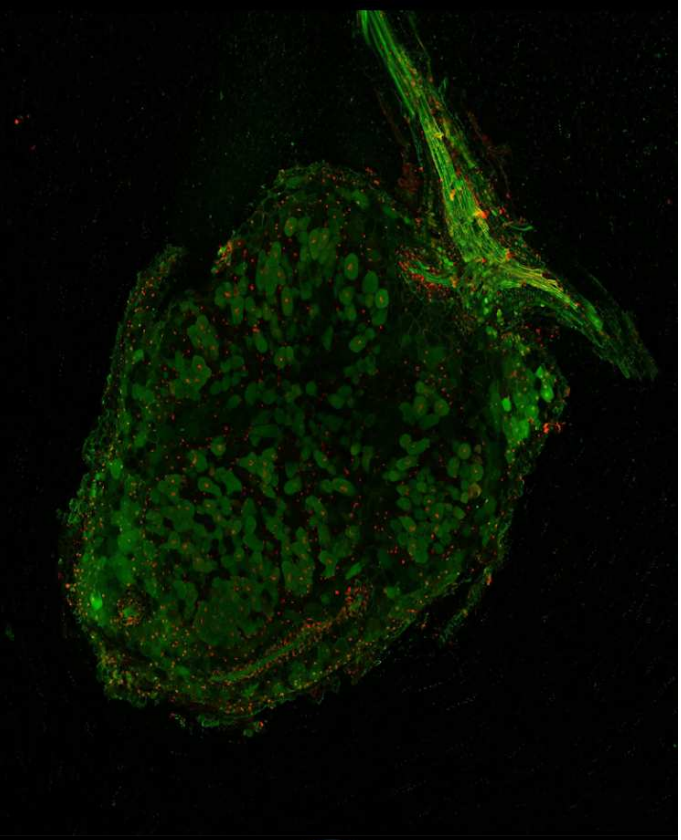
Zhang, X., Jia, X., Yan, L., Wang, J., Kang, X., & Cui, L. (2017). Cyanobacterial Nitrogen Fixation Influences the Nitrogen Removal Efficiency in a Constructed Wetland. *Water*, *9*(11), Article 11. https://doi.org/10.3390/w9110865

Zhou, P., Silverstein, K. A., Gao, L., Walton, J. D., Nallu, S., Guhlin, J., & Young, N. D. (2013). Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinformatics*, *14*, 335. https://doi.org/10.1186/1471-2105-14-335

Zhu, H., Ren, X., Yang, X., Liang, X., Liu, A., & Wu, G. (2022). Fe-based catalysts for nitrogen reduction toward ammonia electrosynthesis under ambient conditions. *SusMat*, *2*(3), 214–242. https://doi.org/10.1002/sus2.70

Zorin, E. A., Kliukova, M. S., Afonin, A. M., Gribchenko, E. S., Gordon, M. L., Sulima, A. S., Zhernakov, A. I., Kulaeva, O. A., Romanyuk, D. A., Kusakin, P. G., Tsyganova, A. V., Tsyganov, V. E., Tikhonovich, I. A., & Zhukov, V. A. (2022). A variable gene family encoding nodule-specific cysteine-rich peptides in pea (Pisum sativum L.). *Frontiers in Plant Science*, *13*. https://www.frontiersin.org/articles/10.3389/fpls.2022.884726





ANNEXES

PNAS

Main Manuscript for

Hundreds of antimicrobial peptides create a selective barrier for insect gut symbionts

Joy Lachat[a,1], Gaëlle Lextrait[a,1], Romain Jouan[a,1], Amira Boukherissa[a], Aya Yokota[a], Seonghan Jang[b,c,2], Kota Ishigami[b,c], Ryo Futahashi[d], Raynald Cossard[a], Delphine Naquin[a], Vlad Costache[e], Luis Augusto[a], Pierre Tissières[a], Emanuele G. Biondi[a], Benoît Alunni[a,3], Tatiana Timchenko[a], Tsubasa Ohbayashi[a,4], Yoshitomo Kikuchi[b,c,*], Peter Mergaert[a,*]

[a]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC); Gif-sur-Yvette, 91198, France.
[b]Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Hokkaido Center; Sapporo, 062-8517, Japan.
[c]Unit of Applied Biological Chemistry, Graduate School of Agriculture, Hokkaido University, 060-8589 Sapporo, Japan.
[d]National Institute of Advanced Industrial Science and Technology (AIST); Tsukuba, 305-8566, Japan.
[e]MIMA2 Imaging Core Facility, Microscopie et Imagerie des Microorganismes, Animaux et Aliments, INRAE; Jouy-en-Josas, 78352, France.

[1]These authors contributed equally to this work
[2]Present address: Infectious Disease Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, South Korea.
[3]Present address: Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay; 78000 Versailles, France.
[4]Present address: Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization (NARO); Tsukuba, 305-8604, Japan.

*Corresponding authors: Peter Mergaert and Yoshitomo Kikuchi

Email: peter.mergaert@i2bc.paris-saclay.fr; y-kikuchi@aist.go.jp

Author Contributions: P.M. and Y.K. designed the study, planned the experiments and supervised the project. T.O., R.F, A.B. and B.A. performed transcriptome analysis. J.L., R.J., D.N. performed Tn-seq. J.L., R.J. and T.T. made mutants. P.M. and E.B. performed *in vitro* peptide activity assays. S.J. and K.I. provided strains. V.C. performed SEM. L.A. and P.T. performed LPS characterization. J.L., G.L., A.Y., R.C and T.O. performed insect experiments. J.L., G.L., R.J, A.B., T.O., Y.K, and P.M. analyzed data. P.M. wrote the manuscript with input from Y.K. All authors provided critical feedback and helped to shape the manuscript.

Competing Interest Statement: Authors declare no competing interests.

Classification: Biological Sciences, Microbiology

Keywords: Gut microbiota biogeography, antimicrobial peptides, insect, resistance

40

41 **This PDF file includes:**

42 Main Text
43 Figures 1 to 5

44

Abstract (max 250 words)

The spatial organization of gut microbiota is crucial for the functioning of the gut ecosystem, although the mechanisms that organize gut bacterial communities in microhabitats are only partially understood. The gut of the insect *Riptortus pedestris* has a characteristic microbiota biogeography with a multispecies community in the anterior midgut and a mono-specific bacterial population in the posterior midgut. We show that the posterior midgut region produces massively hundreds of specific antimicrobial peptides (AMPs), the Crypt-specific Cysteine-Rich peptides (CCRs) that have membrane-damaging antimicrobial activity against diverse bacteria but posterior midgut symbionts have elevated resistance. We determined by transposon-sequencing the genetic repertoire in the symbiont *Caballeronia insecticola* to manage CCR stress, identifying different independent pathways, including novel AMP-resistance pathways unrelated to known membrane homeostasis functions as well as cell envelope functions. Mutants in the corresponding genes have reduced capacity to colonize the posterior midgut, demonstrating that CCRs create a selective barrier and resistance is crucial in gut symbionts. Moreover, once established in the gut, the bacteria differentiate into a CCR-sensitive state, suggesting a second function of the CCR peptide arsenal in protecting the gut epithelia or mediating metabolic exchanges between the host and the gut symbionts. Our study highlights the evolution of an extreme diverse AMP family that likely contributes to establish and control the gut microbiota.

Significance Statement (max 120 words)

The microbioata is usually not homogeneously dispersed in the animal gut but spatially structured in microenvironments. The microbiota in the gut of the bean bug *Riptortus pedestris* displays a sharp divide between the anterior and posterior midgut with a multispecies bacterial community in the anterior region and a specific, mono-species *Caballeronia* symbiont population in the posterior region. We found that this insect deploys in the midgut an arsenal of several hundreds of antimicrobial peptides that creates a selective environment restricting the type of bacteria from the anterior midgut microbiota that have a chance to establish in the posterior midgut. This finding highlights a mechanism that could contribute in the construction of an exclusive niche for beneficial gut symbionts.

Main Text

Introduction

The animal gut is colonized by bacterial communities, which provide essential functions to the host (1, 2). The phylotype richness and total abundance of this gut microbiota varies strongly among the animals from low to extraordinarily high (1). Moreover, in animals ranging from humans to insects, gut microbiota do not constitute a homogeneous mixture but are spatially organized and form discrete bacterial communities located in specific microhabitats along the longitudinal and transverse axes of the gut (3-6). How this microbial biogeography is established is only partially understood but is potentially correlated with physical barriers such as mucus, peritrophic membrane and crypts, gradients of chemical parameters such as pH or oxygen levels, bacteriophages and nutrient availability as well as host immune effectors. Among the latter are antimicrobial peptides (AMPs), which are secreted in the gut lumen and come in contact with the microbiota (7-10). AMPs contribute to establish an epithelia-microbiota equilibrium along the transverse axis of the gut by regulating the species composition and location of the microbiota according to the resistance and sensitivity patterns of its members (10). Thus, gut commensals are expected to be resilient to AMPs (11, 12) but how they adapt and how important this adaptation is for colonization of their specific niche within the gut remains largely unexplored. Moreover, it is not known if AMPs exert control on the spatial organization of microbiota along the longitudinal gut axis.

The bean bug *Riptortus pedestris* has a particular midgut organization, associated with a simple microbiota displaying a characteristic biogeography. The midgut has four morphologically and functionally distinct compartments, labelled M1 to M4. The anterior M1 to M3 regions are involved in food digestion and have a variable and transient microbiota, which is ingested through feeding. The posterior M4 region on the other hand, composed of two rows of crypts branched on a central tract, does not contribute to food digestion and is associated with a stable, (nearly) mono-specific and high-abundant microbiota that is also acquired from the environment and sorted out from the M3 microbiota (13, 14). The M4 bacteria are very specific, belonging to the *Caballeronia* genus and are mostly present as a single colonizing species, established through a multifaceted selection process. A sorting organ located at the entry of the M4 region winnows out the M3 microbiota allowing only a subset of species to enter the M4 (15). After a successful initial passage of bacteria through the sorting organ and infection of the M4 region, secondary infections are inhibited by closure of the sorting organ (16). The infecting bacteria induce in the M4 crypts developmental processes, including oxygenation by tracheal formation (17) and the maturation of the crypts by intestinal stem cell stimulation and apoptosis inhibition that creates the luminal space in the crypts for bacterial colonization (18). Finally, microbe-microbe competition within the crypts results in the elimination of the least adapted strains and the dominance of a single strain in the M4 region (19). Among the bean bug colonizers, *Caballeronia insecticola* has emerged as a model species (20). We took advantage of this simplified gut-microbe interaction model to explore if together with the already known mechanisms, AMP challenge contributes to create the gut biogeography in *R. pedestris* and if AMP resistance in *C. insecticola* is crucial for M4 crypt colonization.

Results

The *Riptortus pedestris* midgut expresses hundreds of AMP-like genes

A preliminary transcriptome analysis of the M4 midgut region has identified a novel class of secretory peptides, which we call the Crypt-specific Cysteine-Rich peptides or CCRs (21, 22). In order to define the expression pattern of *CCR* genes, the transcriptome was determined by RNA-seq in midgut regions of insects that were reared for different times in the presence or absence of the *C. insecticola* gut symbiont (Fig. 1A). The pooled sequencing reads were assembled in a set of unique transcripts and encoded proteins. Hidden Markov Models based on the previously identified CCR sequences were used to identify in the newly generated transcriptome the complete set of CCR sequences. This analysis revealed 310 *CCR* transcripts (SI Appendix, Data S1). Together, these transcripts encode 217 distinct CCR peptides derived from 126 putative genes. Closely related transcripts and peptide variants could arise from recently duplicated genes, alternative splicing of gene transcripts or allelic variation present in the rearing population, although the latter is expected to be low since it is an inbred line derived from a single pair. The CCR peptides do not show high similarity apart from a pattern of conserved cysteine residues (Fig. 1B). Despite their sequence divergence, AlphaFold2 predicted similar folds for tested CCR peptides, consisting of three pairs of β-sheets that are probably connected by cystine bridges (Fig. 1C). Differential expression analysis revealed that the majority of the *CCR* genes are most strongly expressed in the midguts of symbiotic insects (Fig. 1D and SI Appendix, Data S1). Subsets of genes were specific for the M3, M4B and the majority for the M4 region carrying the *C. insecticola* bacteria, suggesting that the encoded peptides target the symbionts. Moreover, the *CCRs* are among the most strongly expressed transcripts in the overall transcriptome (Fig. 1E), suggesting a primordial role of the peptides in the midgut. The *CCR* genes did not exhibit similarity to known sequences of other organisms. However the taxonomically restricted nature of the genes as well as the structure of the CCRs, being small, secreted and characterized by conserved cysteine residues, are features shared with AMP gene families (10) and AMP prediction tools confirmed this presumption (Fig. 1C and SI Appendix, Data S1). Whole mount *in situ* hybridization with the infected-M4-specific gene *CCR0043* showed that the gene is expressed uniformly by the epithelial cells in all M4 crypts (Fig.

1F and SI Appendix, Fig. S1). This pattern contrasts with the mammalian small intestine where specialized cell types at the base of crypts express AMP genes (23).

CCRs have antibacterial activity but gut colonizers are resistant

We selected CCRs for chemical synthesis on the basis of a consistent prediction of AMP activity, an independently confirmed transcript sequence by cloned cDNA sequencing (21), taking into account the diversity of expression patterns, including peptides expressed in apo and/or sym insects and in the M4B and/or M4, with high or medium expression levels, and favouring smaller peptides to increase feasibility of successful peptide synthesis (synthesis attempts for several initially selected peptides failed) (Fig. 1C and SI Appendix, Table S1). A total of seven CCRs, together with thanatin and riptocin, two known innate immunity-related AMPs of *R. pedestris* (24), LL37 and NCR335, from mammal and plant origin respectively (25, 26) and bacterial polymyxin B (PMB), were tested for growth inhibiting activity against a panel of taxonomically diverse bacterial species consisting in *Bacillus subtillis*, *Sinorhizobium meliloti*, *Paraburkholderia fungorum* and *C. insecticola*. The first two species are typical, well-studied soil bacteria known to be unable to colonize the *R. pedestris* M4 crypts while the latter two can efficiently proliferate in the crypts (19). In agreement with the bioinformatics predictions, the CCRs had growth inhibiting activity against *B. subtilis* and *S. meliloti* although with variable strengths (Fig. 2A, B). On the other hand, the two species, *P. fungorum* and *C. insecticola*, that are able to colonize the gut crypts, are not or only weakly affected by the tested CCRs (Fig. 2A, B). This pattern of sensitivity/resistance to CCRs matches with the response of these species to PMB and in part to the other tested peptides.

CCRs have membrane-damaging bactericidal activity

CFU counting showed the bacterial reduction from 10^7 CFU to no colonies after treatment of sensitive *S. meliloti* with the CCR1659 peptide for a few hours, indicating that the growth inhibition results from a bactericidal activity, similarly as for PMB (Fig. 3A). The bactericidal activity of CCR1659 was abolished by prior Proteinase K treatment of the peptide and inhibited by the presence of the divalent cations Ca^{2+} and Mg^{2+}, which interfere with the electrostatic interaction of AMPs with negatively charged membrane lipids and diminish the activity of membrane-targeting AMPs (27) (Fig. 3A; SI Appendix, Fig. S2). To acquire insight in the killing mode of CCR1659, we tested the hypothesis that the peptide disrupt bacterial membranes, like PMB and the other tested AMPs do (28-30). Outer and inner membrane integrities in *S. meliloti* were consecutively damaged by both CCR1659 and PMB treatment, as measured respectively by 1-N-PhenylNaphthylamine (NPN) and Propidium Iodide (PI) uptake leading to enhanced fluorescence (Fig. 3B). In agreement with the membrane disruption, fluorescence microscopy showed that FITC-modified CCR1659 labelled the envelope of *S. meliloti* cells in a similar way as polylysine-FITC, which is a polycation known to interact with negatively charged membranes of bacteria (31, 32) (Fig. 3C). Binding of CCR1659 to the envelope suggests that its killing efficiency depends on the strength of envelope binding. To test this assumption, we measured with flow cytometry the binding level of CCR1659-FITC and polylysine-FITC to the above panel of species. Strikingly, CCR1659-sensitive *S. meliloti* and *B. subtilis* were strongly labeled with these two molecules while resistant *C. insecticola* and *P. fungorum* only weakly (Fig. 3D). Thus, the level of binding to cells is correlated with the susceptibility/resistance pattern. Scanning electron microscopy (SEM) of CCR1659-treated *S. meliloti* cells further confirmed the membrane-perturbing activity of the peptide that provoked the formation of fibrous materials from damaged cells similarly as PMB (Fig. 3E; SI Appendix, Fig. S3) and similarly as reported for bacteria and yeasts treated with other types of membrane-disrupting AMPs (33-38). Together, this data reveal that the M4 symbiotic region of the gut produces a remarkably large arsenal of CCR peptides with membrane-damaging AMP activity.

The *Caballeronia insecticola* genetic repertoire determining AMP resistance

Species that colonize the midgut display a high level of resistance to CCRs and other AMPs suggesting that resistance is a prerequisite for efficient gut colonization. To test this hypothesis, we aimed to identify the resistance determinants in *C. insecticola* and assess if they control gut colonization. A transposon mutant library (39) was used to perform a Tn-seq screen with PMB, since PMB has a similar membrane action as CCR peptides and is commercially accessible in sufficient quantities for Tn-seq experiments. The screen, performed with three sub-lethal PMB concentrations, resulted in 54 genes whose mutation provoked a fitness defect with the highest concentration. With the lower PMB concentrations, subsets of these genes were identified suggesting a multifactorial resistance with some mechanisms contributing more strongly than others (Fig. 4A, B; SI Appendix, Fig. S4 and Data S2). In agreement with the membrane-targeting mode of action of PMB, the majority of fitness genes are involved in the generation of bacterial envelope components, including LPS, peptidoglycan, phospholipids, hopanoids and membrane protein machineries. In order to validate the Tn-seq results, we constructed insertion and deletion mutants in 11 genes selected among the 54 PMB fitness genes. These genes are predicted to be involved in the biosynthesis of the LPS core (*dedA*, *waaC* and *waaF*) (40-42), LPS O-antigen (*wbiF*, *wbiG*, *wbiI*, *wzm* and *rfbA*) (43), peptidoglycan (*dedA*), membrane protein machineries (*tolB* and *tolQ*) (44, 45), in addition to a gene (*tpr*) encoding a tetratricopeptide repeat protein of unknown function. Complementing strains were constructed for some of the mutants. Sensitivity assays with PMB and colistin (COL), another polymyxin-family AMP, confirmed that each mutant had an 8- to 32-fold increased sensitivity compared to the WT (Fig. 4C) while the complemented mutants were restored to WT-levels (SI Appendix, Fig. S5). Thus, the Tn-seq analysis correctly identified genetic determinants for PMB resistance in *C. insecticola*.

In line with the sensitivity of the mutants to PMB and the membrane-attacking properties of CCRs, we found that all mutants were more sensitive than WT for at least one of the tested CCR peptides and the other available AMPs (Fig. 4C, D). The *tolB* and *tolQ* mutants were the least sensitive and displayed only a slight difference compared to WT for all tested peptides. The *dedA* and *tpr* mutants were strongly affected by the CCR1659 peptide (Fig. 4D) and moderately by the other tested peptides. The mutants *wzm*, *wbiF*, *wbiG*, *wbiI* and *rfbA* were sensitive to several of the tested CCRs although in many cases, enhanced sensitivity was not resulting in a complete growth inhibition but in a retarded and lesser growth compared to untreated control and the WT grown with the same peptide concentration. The *waaC* and *waaF* mutants were the most strongly affected, being more sensitive than WT to all tested peptides and at higher peptide concentrations, their growth was completely blocked (Fig. 4D). Taken together, the *C. insecticola* genes that were revealed by the PMB Tn-seq screen, contribute also to resistance towards other membrane-attacking AMPs, including the CCRs.

Different pathways contribute to AMP resistance in *Caballeronia insecticola*

Because the tested AMPs interfere with bacterial membrane function, we characterized the cell envelope of the mutants. Since some of the mutated genes are known or suspected to be involved in LPS biosynthesis, we analyzed the LPS structure of all mutants by PAGE profiling and by mass spectrometry analysis of their lipid A moiety, which is proposed to be a direct target of PMB (28, 29) (Fig. 5A, B). The *tpr*, *dedA* and *tolQ* mutants had a PAGE LPS profile that was indistinguishable from the WT. The *wzm*, *rfbA*, *wbiF*, *wbiG* and *wbiI* mutants produced a similar LPS that lacked the O-antigen but had a lipid A/core oligosaccharide moiety that was indistinguishable from WT while the *waaC* and *waaF* mutants had an altered lipid A/core moiety, in agreement with the predicted heptosyl-transferase activity of the encoded enzymes that perform the first steps of the core oligosaccharide synthesis. Mass spectrometry analysis of the lipid A moieties suggested that none

of the mutants had an altered lipid A structure and notably, that all mutants produced lipid A carrying the 4-amino-4-deoxy-L arabinose (Ara4N) modification that is known to confer PMB resistance in related bacterial species (41, 42, 46) (Fig. 5B; SI Appendix, Fig. S6).

We assessed the steady-state outer membrane integrity of the mutants by NPN labeling and sensitivity to detergents (SI Appendix, Fig. S7). The *waaC* and *waaF* mutants had a higher NPN-derived fluorescence and slightly higher sensitivity to the non-ionic detergent Triton X100 and the cationic detergent CTAB than the WT, while the other mutants were similar to WT. The *tolB* and *tolQ* mutants on the other hand were more sensitive to the anionic detergent SDS than to other tested strains. Overall, this indicates that although the outer membrane in some mutants has a reduced robustness, the AMP sensitivity of the mutants is not a direct consequence of a generic membrane instability but of the deficiency of specific resistance mechanisms.

The capacity of the bacterial envelope to bind membrane-disrupting AMPs is a parameter influencing AMP sensitivity. The *waaC* and *wbiF* LPS mutants showed indeed a strong labeling of their envelope with CCR1659-FITC, contrary to the WT that did not show any labelling (Fig. 5C). However, the *tpr* mutant was also not labeled. Therefore, we quantified the relative capacity of the envelope of all the mutants to bind membrane-disrupting AMPs by labeling the cells with the fluorescent polylysine-FITC peptide or CCR1659-FITC, followed by flow cytometry analysis (Fig. 5D). All the mutants with altered LPS (*waaC*, *waaF*, *wzm*, *rfbA*, *wbiF*, *wbiG* and *wbiI*) had a strongly enhanced labeling with both peptides indicating a more accessible cell surface for AMP binding. However, the *dedA* and *tpr* mutants displayed a peptide labeling that was identical to the WT while the *tolB* and *tolQ* mutants were even labelled less intensively. Thus, the LPS mutants might be more sensitive to the AMPs because of the higher accessibility of their membranes for interactions with AMPs but the sensitivity of the *dedA*, *tpr* and *tolBQ* mutants has to be explained by a different mechanism. Interestingly, crypt-colonizing *C. insecticola* bacteria have lost their O-antigen after establishing in the crypts (24) and thus have an LPS that is similar to the LPS of the *wzm*, *rfbA*, *wbiF*, *wbiG* and *wbiI* mutants. In agreement, bacteria isolated from the crypts are hypersensitive to PMB and the CCR1659 peptide (Fig. 4C, D) and they strongly bind polylysine-FITC and CCR1659-FITC (Fig. 5D).

To confirm that the set of mutants are affected in different pathways for AMP resistance, we created the *waaC*/*tpr*, *waaC*/*dedA* and *waaC*/*wbiF* double mutants. We reasoned that if genes are part of the same pathway, double mutants should not show an additive phenotype compared to the single mutants, while in case genes are in separate pathways, double mutants might display a more severe phenotype than single mutants. We found that the three double mutants were more sensitive than the corresponding single mutants to PMB and CCR1659 and bound more CCR1659-FITC (SI Appendix, Fig. S8), suggesting that indeed "*waaC* and *tpr*" or "*waaC* and *dedA*" or "*waaC* and *wbiF*" define different pathways to PMB resistance. The synthetic phenotype of the *waaC*/*wbiF* mutant further suggest that the LPS core and the O-antigen constitute two distinct barriers for AMPs to reach the membrane.

SEM of untreated WT and *tpr*, *dedA*, *tolB*, *waaC* and *wzm* mutants showed that the mutants affect the bacterial envelope in various ways (SI Appendix, Fig. S9). SEM of CCR1659-treated cells reveals that the response to the peptide in the *waaC* and *tpr* mutant is markedly different. In the *waaC* mutant, very strong membrane distortions are visible and frequent cell lysis, indicated by the cellular material released from cells. The *tpr* mutant on the other hand shows only minor modifications on the cell surface, similar to WT, although infrequent release of large amounts of cellular material was also observed (Fig. 5E). Collectively, the properties of the single and double mutants suggest that in *C. insecticola* different mechanisms contribute to AMP resistance.

AMP resistance in *Caballeronia insecticola* is crucial for midgut colonization

Since the midgut crypts are the site of intensive AMP production, we next analyzed the capacity of the AMP sensitivity mutants to colonize the M4 midgut region of the *R. pedestris* midgut. As a preliminary test and to exclude that gut colonization phenotypes can be attributed to trivial reasons, we confirmed that each mutant has similar growth patterns as WT (SI Appendix, Fig. S10A) and is motile (SI Appendix, Fig. S10B) since motility is crucial for colonization of the M4 crypts (14). Analysis at 5 days post infection (dpi) of second instar nymphs showed that the 11 mutants had the capacity to colonize the crypts although they were to various extends less efficient than the WT. The WT had a 100% efficiency (n=10) and the number of bacteria per gut was consistently high ($>10^7$ genome copies per gut). In contrast, the mutants displayed a large variability in colonization level between insect individuals, ranging from a wild-type colonization level for some individuals to a failure to establish in the crypts in other individuals (Fig. 6A). The *waaC* and *tpr* mutants were particularly affected in agreement with their strong AMP sensitivity. This intriguing probabilistic colonization of the gut by the mutants is reminiscent to stochastic colonization of the *Drosophila* gut by underperforming *Lactobacillus plantarum* strains while a strong colonizer strain had a 100% efficiency (47).

Next, we evaluated the fitness of the mutants in M4 colonization when they were in competition with WT. Insects were infected with fifty-fifty mixtures of RFP-marked WT and one of the mutants (or WT as a control) that were marked with GFP. The outcome of the competitions was analyzed at 5 dpi by fluorescence microscopy of dissected M4 midguts and flow cytometry quantification of their bacterial content (Fig. 6B). In the control competition, RFP- and GFP-marked WT were kept in balance in the M4 crypts. However, in competitions with the mutants, the WT nearly completely outcompeted each of them, confirming their reduced colonization capacity.

Finally, we also tested if the mutants have maintained or lost the capacity to outcompete a less efficient crypt colonizing species. We previously showed that *P. fungorum* can efficiently colonize the M4 crypts in the absence of competing strains but that it is outcompeted by *C. insecticola* when both strains are co-infecting the *R. pedestris* midgut (19). Here, the outcompetition in the M4 crypts of *P. fungorum* by *C. insecticola* WT in co-infection experiments was confirmed while *wzm*, *waaC*, *tolB*, *tpr* and *dedA* mutants were significantly less efficient in outcompeting *P. fungorum* (Fig. 6C). Thus, high AMP resistance in *C. insecticola* is an important factor contributing to the efficiency of this strain in occupying the *R. pedestris* gut.

Discussion

The microbiota biogeography in the *R. pedestris* midgut shows a sharp divide between the anterior midgut, which has a highly variable, diverse and relatively low abundant microbiota, and the posterior midgut region that carries in striking contrast a dense mono-specific bacterial population that is strictly a *Caballeronia* species. The two principal findings from this work are that this posterior midgut region and the immediately adjacent anterior region is a highly challenging environment for bacteria because of the abundant presence of symbiosis-specific, membrane-damaging antimicrobial CCRs (Figs. 1-3) and that resistance to these AMPs is crucial for bacteria to colonize the crypts in the posterior midgut (Figs. 4-6). Thus, we propose that the CCRs are new players, acting together with previously identified sorting mechanisms (14, 16-19), in the creation of the biogeography by eliminating sensitive bacteria. A prediction that follows from this hypothesis is that the prevention of CCR production in the gut (for example by the knock-down of a global regulator of *CCR* gene expression) would alleviate the strict selectivity for *Caballeronia* in the M4 crypts.

The expression of the majority of the *CCR* genes is correlated with crypt colonization because they are specifically expressed in the M4 crypt region of the midgut and they are frequently induced by bacterial colonization of the crypts (Fig. 1). A few of them are also expressed in the upstream M3 midgut region, where they still may have a function related to crypt colonization, for example by preselecting bacterial species. Overall, their gene expression pattern, combined with their

secretory nature suggesting that they are released in the lumen of the midgut, is consistent with a function of the CCRs in interacting with the bacterial community during midgut infection as well as during M4 colonization.

Our analyses demonstrated that CCR peptides act through membrane interaction and damage (Fig. 3), similarly to most AMPs produced by eukaryotic organisms (30, 48). AMPs damage membranes by first interacting with negative charges exposed on the membrane. In many Gram-negative bacteria, the negative charges carried by phosphate groups on the lipid A moiety of LPS are particularly important for this electrostatic interaction (29, 30). However, in *C. insecticola*, including in the AMP-sensitive mutants, these lipid A charges are converted into positive charges through the Ara4N modifications and therefore, the lipid A is likely not the target of AMPs in *C. insecticola*. We propose that the O-antigen and core oligosaccharide of LPS form a safeguard around the cell that limits the access of AMPs to their targets in the membrane. This hypothesis is consistent with the enhanced sensitivity and peptide-binding of mutants without O-antigen or LPS core (Fig. 5). The direct targets of the AMPs are presently unknown but could be revealed by the analysis of the other genetic determinants of AMP resistance in *C. insecticola* identified here, including *tpr*, *dedA*, *tol-pal*, *tamAB*, *rpoE*, and hopanoid and phospholipid biosynthesis genes.

We conclude from our infection experiments that the reduced resilience to CCRs of *C. insecticola* mutants in different resistance mechanisms makes them less apt to colonize the midgut crypts (Fig. 6). This correlates with the inability of strongly sensitive bacterial species to colonize the midgut (Fig. 2) (19). Presumably, AMP-resistance is critical during the initial infection stages, when a few hundred cells enter into the crypt region and this founder population subsequently multiplies rapidly, in two to three days, to a crypt-space-filling population of about 10^7-10^8 bacteria (14, 16). The surprisingly large diversity of CCR peptides, several of them already expressed in the M3 and M4 before the microbiota establishment, could be an adaptation to create a selective environment that restricts the type of bacteria from the anterior midgut microbiota that have a chance to establish in the M4 crypts and that favors optimal beneficial *Caballeronia* strains. Such a molecular filter of bacteria could arise from additive, synergistic or specific antimicrobial activities of different CCR peptides towards distinct bacteria. Indeed, the tested CCRs have variable antimicrobial efficiency against different bacterial species and *C. insecticola* mutants. Recent insights from *Drosophila* and other models have changed the previous view on AMPs as generic, non-specific antimicrobials by the demonstration that they can display a degree of specificity and synergism. Accordingly, AMP repertoires in organisms dynamically evolve according to the diversity of microbes encountered in the natural environment (30, 48, 49). The hundreds of diverse CCR peptides might be an extreme example of such an evolutionary process.

On the other hand, once established in the M4 crypts, *C. insecticola* loses its O-antigen by an unknown mechanism (24), which renders them sensitive to the CCR peptides (Fig. 4C, D). This suggests a second function of the CCR peptide arsenal - in particular for those peptides encoded by the late-expressed genes - that could be related to the protection of the crypt epithelia and prevention of the bacteria breaching these epithelia. Indeed, in *R. pedestris* the crypt epithelium lacks mucus or peritrophic protective layers and is therefore in direct contact with the microbiota (17). Additionally, the membrane fragilization of the crypt-colonizing bacteria by the CCRs could facilitate the retrieval of nutrients from the bacteria (50), suggesting that the insect tames the gut symbionts with the CCRs. Although these additional functions should be confirmed in future studies, they may also be related to the CCR diversification in *R. pedestris*.

Materials and Methods

Detailed protocols for all used procedures are available in SI Appendix. Bacterial strains and materials used are listed in SI Appendix, Table S1. Tn-seq screening with PMB was done with an available *Himar1* transposon mutant Tn-seq library (39). Plasmid constructs for mutagenesis or

mutant complementation were obtained by gene synthesis or Gibson cloning. Plasmids were introduced in *C. insecticola* by mating and deletion mutants were obtained by double homologous recombination. Strains were tagged with fluorescent proteins using mutagenesis with modified Tn7 transposons. Antimicrobial peptide activity and detergent sensitivity assays were performed in 96-well plates and determining growth curves in the presence of test peptide dilution series or by cfu counting after peptide exposure during variable times. Membrane interactions of CCR peptides were determined by NPN and PI uptake assays, fluorescence microscopy, flow cytometry and SEM. LPS was obtained by phenol extraction. Total LPS was analyzed by SDS-PAGE and the lipid A fraction, obtained by triethylamine-citrate treatment, was analyzed by matrix-assisted laser desorption ionization–time of flight mass spectrometer.

The rearing conditions for *R. pedestris* and the procedures for colonization assays were done as before (13, 39) and are detailed in SI Appendix. Colonization efficiency by tested bacterial strains were determined by light and fluorescence microscopy, qPCR and flow cytometry. Transcriptome analysis of the midgut of colonized and aposymbiotic insects was done by RNA-seq and Illumina sequencing. Whole-mount *in situ* hybridization on the *R. pedestris* midgut was performed with digoxigenin (DIG)-labeled *CCR043* cRNA probe. Bioinformatic procedures for data analysis are available in SI Appendix.

Data availability

RNA-seq sequencing data are available in the Sequence Read Archive (SRA), BioProject accession no. PRJNA1006624. The *de novo* assembled transcriptome was deposited in the Transcriptome Shotgun Assembly (TSA), BioProject accession no. PRJNA1006624. Tn-seq sequencing data were deposited in SRA, BioProject accession no. PRJNA890438.

Acknowledgments

We are grateful to Olga Soutourina (University Paris-Saclay, France) and Yu Matsuura (University of the Ryukyus, Japan) for critical reading of the manuscript and constructive comments. This work benefited from financial support by Saclay Plant Sciences-SPS, by the ANR grant ANR-19-CE20-0007, and by a JSPS-CNRS Bilateral Open Partnership Joint Research Project (18KK0211) and a CNRS International Research Project to Y.K. and P.M. Y.K. was supported by the MEXT KAKENHI (21K18241, 22H05068, 22B303). J.L., G.L. and R.J. were supported by Ph.D. fellowships from the French Ministry of Higher Education, Research, and Innovation and A.B. benefited from a PhD contract in the frame of the CNRS 80|PRIME – 2021 program. This work was supported by JSPS Research Fellowships for Young Scientist to S.J. (21F21090), K.I. (22KJ0057) and T.O (14J03996, 20170267 and 19J01106). Tn-seq sequencing and data treatment were performed by the I2BC high-throughput sequencing facility, supported by France Génomique (funded by the French National Program Investissement d'Avenir ANR-10-INBS-09). This work has benefited from the facilities and expertise of MIMA2 (Université Paris-Saclay, INRAE, AgroParisTech, 78350, Jouy-en-Josas, France).

References

1. N. A. Moran, H. Ochman, T. J. Hammer, Evolutionary and ecological consequences of gut microbial communities. *Annual Review of Ecology, Evolution, and Systematics* 50, 451-475 (2019). doi: 10.1146/annurev-ecolsys-110617-062453.

2. J. C. Clemente, L. K. Ursell, L. W. Parfrey, R. Knight, The impact of the gut microbiota on human health: an integrative view. *Cell* 148, 1258-1270 (2012). doi: 10.1016/j.cell.2012.01.035.
3. G. McCallum, C. Tropini, The gut microbiota and its biogeography. *Nature Reviews Microbiology* (2023). doi: 10.1038/s41579-023-00969-0.
4. G. P. Donaldson, S. M. Lee, S. K. Mazmanian, Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology* 14, 20-32 (2016). doi: 10.1038/nrmicro3552.
5. Z. Yao, Z. Cai, Q. Ma, S. Bai, Y. Wang, P. Zhang, Q. Guo, J. Gu, B. Lemaitre, H. Zhang, Compartmentalized PGRP expression along the dipteran *Bactrocera dorsalis* gut forms a zone of protection for symbiotic bacteria. *Cell Reports* 41, 111523 2022). doi: 10.1016/j.celrep.2022.111523.
6. R. Dodge, E.W. Jones, H. Zhu, B. Obadia, D.J. Martinez, C. Wang, A. Aranda-Díaz, K. Aumiller, Z. Liu, M. Voltolini, E. L. Brodie, K. C. Huang, J. M. Carlson, D. A. Sivak, A. C. Spradling, W. B. Ludington, A symbiotic physical niche in *Drosophila melanogaster* regulates stable association of a multi-species gut microbiota. *Nature Communications* 14, 1557 (2023). doi: 10.1038/s41467-023-36942-x.
7. Z. Hu, C. Zhang, L. Sifuentes-Dominguez, C. M. Zarek, D. C. Propheter, Z. Kuang, Y. Wang, M. Pendse, K. A. Ruhn, B. Hassell, C. L. Behrendt, B. Zhang, P. Raj, T. A. Harris-Tryon, T. A. Reese, L. V. Hooper, Small proline-rich protein 2A is a gut bactericidal protein deployed during helminth infection. *Science* 374, eabe6723 (2021). doi: 10.1126/science.abe6723.
8. D. Sun, R. Bai, W. Zhou, Z. Yao, Y. Liu, S. Tang, X. Ge, L. Luo, C. Luo, G. F. Hu, J. Sheng, Z. Xu, Angiogenin maintains gut microbe homeostasis by balancing α-Proteobacteria and Lachnospiraceae. *Gut* 70, 666-676 (2021). doi: 10.1136/gutjnl-2019-320135.
9. R. L. Gallo, L. V. Hooper, Epithelial antimicrobial defence of the skin and intestine. *Nature Reviews Immunology* 12, 503-516 (2012). doi: 10.1038/nri3228.
10. P. Mergaert, Role of antimicrobial peptides in controlling symbiotic bacterial populations. *Natural Product Reports* 35, 336-356 (2018). doi: 10.1039/c7np00056a.
11. T. W. Cullen, W. B. Schofield, N. A. Barry, E. E. Putnam, E. A. Rundell, M. S. Trent, P. H. Degnan, C. J. Booth, H. Yu, A. L. Goodman, Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* 347, 170-175 (2015). doi: 10.1126/science.1260580.
12. A. Arias-Rojas, D. Frahm, R. Hurwitz, V. Brinmann, I. Iatsenko, Resistance to host antimicrobial peptides mediates resilience of gut commensals during infection and aging in *Drosophila*. *Proceedings of the National Academy of Sciences USA* 120, e2305649120 (2023). doi: 10.1073/pnas.2305649120.
13. Y. Kikuchi, T. Hosokawa, T. Fukatsu, Insect-microbe mutualism without vertical transmission: a stinkbug acquires a beneficial gut symbiont from the environment every generation. *Applied and Environmental Microbiology* 73, 4308-4316 (2007). doi: 10.1128/AEM.00067-07.
14. T. Ohbayashi, K. Takeshita, W. Kitagawa, N. Nikoh, R. Koga, X. Y. Meng, K. Tago, T. Hori, M. Hayatsu, K. Asano, Y. Kamagata, B. L. Lee, T. Fukatsu, Y. Kikuchi, Insect's intestinal organ for symbiont sorting. *Proceedings of the National Academy of Sciences USA* 112, E5179-5188 (2015). doi: 10.1073/pnas.1511454112.
15. S. Jang, K. Ishigami, P. Mergaert, Y. Kikuchi, Ingested soil bacteria breach gut epithelia and prime systemic immunity in an insect. *Proceedings of the National Academy of Sciences USA* 121, e2315540121 (2024). doi: 10.1073/pnas.2315540121.
16. Y. Kikuchi, T. Ohbayashi, S. Jang, P. Mergaert, *Burkholderia insecticola* triggers midgut closure in the bean bug *Riptortus pedestris* to prevent secondary bacterial infections of midgut crypts. *The ISME Journal* 14, 1627-1638 (2020). doi: 10.1038/s41396-020-0633-3.
17. S. Jang, P. Mergaert, T. Ohbayashi, K. Ishigami, S. Shigenobu, H. Itoh, Y. Kikuchi, Dual oxidase enables insect gut symbiosis by mediating respiratory network formation. *Proc. Natl. Acad. Sci. USA* 118, e2020922118 (2021). doi: 10.1073/pnas.2020922118.
18. S. Jang, Y. Matsuura, K. Ishigami, P. Mergaert, Y. Kikuchi, Symbiont coordinates stem cell proliferation, apoptosis, and morphogenesis of gut symbiotic organ in the stinkbug-

Caballeronia symbiosis. *Frontiers in Physiology* 13, 1071987 (2023). doi: 10.3389/fphys.2022.1071987.
19. H. Itoh, S. Jang, K. Takeshita, T. Ohbayashi, N. Ohnishi, X. Y. Meng, Y. Mitani, Y. Kikuchi, Host-symbiont specificity determined by microbe-microbe competition in an insect gut. *Proceedings of the National Academy of Sciences USA* 116, 22673-22682 (2019). doi: 10.1073/pnas.1912397116.
20. K. Takeshita, H. Tamaki, T. Ohbayashi, X. Y. Meng, T. Sone, Y. Mitani, C. Peeters, Y. Kikuchi, P. Vandamme, *Burkholderia insecticola* sp. nov., a gut symbiotic bacterium of the bean bug *Riptortus pedestris*. *International Journal of Systematic and Evolutionary Microbiology* 68, 2370-2374 (2018). doi: 10.1099/ijsem.0.002848.
21. R. Futahashi, K. Tanaka, M. Tanahashi, N. Nikoh, Y. Kikuchi, B. L. Lee, T. Fukatsu, Gene expression in gut symbiotic organ of stinkbug affected by extracellular bacterial symbiont. *PLoS One* 8, e64557 (2013). doi: 10.1371/journal.pone.0064557.
22. T. Ohbayashi, R. Futahashi, M. Terashima, Q. Barrière, F. Lamouche, K. Takeshita, X. Y. Meng, Y. Mitani, T. Sone, S. Shigenobu, T. Fukatsu, P. Mergaert, Y. Kikuchi, Comparative cytology, physiology and transcriptomics of *Burkholderia insecticola* in symbiosis with the bean bug Riptortus pedestris and in culture. *The ISME Journal* 13, 1469-1483 (2019). doi: 10.1038/s41396-019-0361-8.
23. H. C. Clevers, C. L. Bevins, Paneth cells: maestros of the small intestinal crypts. *Annual Review of Physiology* 75, 289-311 (2013). doi: 10.1146/annurev-physiol-030212-183744.
24. J. K. Kim, D. W. Son, C. H. Kim, J. H. Cho, R. Marchetti, A. Silipo, L. Sturiale, H. Y. Park, Y. R. Huh, H. Nakayama, T. Fukatsu, A. Molinaro, B.L. Lee, Insect gut symbiont susceptibility to host antimicrobial peptides caused by alteration of the bacterial cell envelope. *Journal of Biological Chemistry* 290, 21042-21053 (2015). doi: 10.1074/jbc.M115.651158.
25. M. F. Burton, P. G. Steel, The chemistry and biology of LL-37. *Natural Product Reports* 26, 1572-1584 (2009). doi: 10.1039/b912533g.
26. A. Farkas, G. Maróti, A. Kereszt, É. Kondorosi, Comparative Analysis of the Bacterial Membrane Disruption Effect of Two Natural Plant Antimicrobial Peptides. *Frontiers in Microbiology* 8, 51 (2017). doi: 10.3389/fmicb.2017.00051.
27. S. J. Ko, E. Park, A. Asandei, J. Y. Choi, S. C. Lee, C. H. Seo, T. Luchian, Y. Park, Bee venom-derived antimicrobial peptide melectin has broad-spectrum potency, cell selectivity, and salt-resistant properties. *Scientific Reports* 10, 10145 (2020). doi: 10.1038/s41598-020-66995-7.
28. L. Poirel, A. Jayol, P. Nordmann, Polymyxins: Antibacterial activity, susceptibility testing, and resistance mechanisms encoded by plasmids or chromosomes. *Clinical Microbiology Reviews* 30, 557-596 (2017). doi: 10.1128/CMR.00064-16.
29. C. A. Moubareck, Polymyxins and bacterial membranes: A review of antibacterial activity and mechanisms of resistance. *Membranes* 10, 181 (2020). doi: 10.3390/membranes10080181.
30. B. P. Lazzaro, M. Zasloff, J. Rolff, Antimicrobial peptides: Application informed by evolution. *Science* 368, eaau5480 (2020). doi: 10.1126/science.aau5480.
31. R. Spohn, L. Daruka, V. Lázár, A. Martins, F. Vidovics, G. Grézal, O. Méhi, B. Kintses, M. Számel, P. K. Jangir, B. Csörgő, Á. Györkei, Z. Bódi, A. Faragó, L. Bodai, I. Földesi, D. Kata, G. Maróti, B. Pap, R. Wirth, B. Papp, C. Pál, Integrated evolutionary analysis reveals antimicrobial peptides with limited resistance. *Nature Communications* 10, 4538 (2019). doi: 10.1038/s41467-019-12364-6.
32. F. F. Rossetti, I. Reviakine, G. Csúcs, F. Assi, J. Vörös, M. Textor, Interaction of poly(L-lysine)-g-poly(ethylene glycol) with supported phospholipid bilayers. *Biophysical Journal* 87, 1711-1721 (2004). doi: 10.1529/biophysj.104.041780.
33. B. O. Schroeder, D. Ehmann, J. C. Precht, P. A. Castillo, R. Küchler, J. Berger, M. Schaller, E. F. Stange, J. Wehkamp, Paneth cell α-defensin 6 (HD-6) is an antimicrobial peptide. *Mucosal Immunology* 8, 661-671 (2015). doi: 10.1038/mi.2014.100.
34. N. Shagahi, M. Bhave, E. A. Palombo, A. H. A. Clayton, Revealing the sequence of interactions of PuroA peptide with *Candida albicans* cells by live-cell imaging. *Scientific Reports* 7, 43542 (2017). doi: 10.1038/srep43542.

35. J. Montiel, J. A. Downie, A. Farkas, P. Bihari, R. Herczeg, B. Bálint, P. Mergaert, A. Kereszt, É. Kondorosi, Morphotype of bacteroids in different legumes correlates with the number and type of symbiotic NCR peptides. *Proceedings of the National Academy of Sciences USA* 114, 5041-5046 (2017). doi: 10.1073/pnas.1704217114.
36. A. Farkas, B. Pap, É. Kondorosi, G. Maróti, Antimicrobial activity of NCR plant peptides strongly depends on the test assays. Frontiers in Microbiology (2018). doi: 10.3389/fmicb.2018.02600.
37. H. Zhuo, X. Zhang, M. Li, Q. Zhang, Y. Wang, Antibacterial and anti-inflammatory properties of a novel antimicrobial peptide derived from LL-37. *Antibiotics* 11, 754 (2022). doi: 10.3390/antibiotics11060754.
38. J. Shi, C. Chen, D. Wang, Z. Wang, Y. Liu, The antimicrobial peptide LI14 combats multidrug-resistant bacterial infections. *Communications biology* 5, 926 (2022). doi: 10.1038/s42003-022-03899-4.
39. R. Jouan, G. Lextrait, J. Lachat, A. Yokota, R. Cossard, D. Naquin, T. Timtchenko, Y. Kikuchi, T. Ohbayashi, P. Mergaert, Transposon sequencing reveals the essential gene set and genes enabling gut symbiosis in the insect symbiont *Caballeronia insecticola*. *ISME Communications* (2024). doi: 10.1093/ismeco/ycad001.
40. J. K. Kim, H. A. Jang, M. S. Kim, J. H. Cho, J. Lee, F. Di Lorenzo, L. Sturiale, A. Silipo, A. Molinaro, B. L. Lee, The lipopolysaccharide core oligosaccharide of *Burkholderia* plays a critical role in maintaining a proper gut symbiosis with the bean bug *Riptortus pedestris*. *Journal of Biological Chemistry* 292, 19226-19237 (2017). doi: 10.1074/jbc.M117.813832.
41. P. R. Panta, S. Kumar, C. F. Stafford, C. E. Billiot, M. V. Douglass, C. M. Herrera, M. S. Trent, W. T. Doerrler, A DedA family membrane protein is required for *Burkholderia thailandensis* colistin resistance. *Frontiers in Microbiology* 10, 2532 (2019). doi: 10.3389/fmicb.2019.02532.
42. P. R. Panta, W. T. Doerrler, A link between pH homeostasis and colistin resistance in bacteria. *Scientific Reports* 11, 13230 (2021). doi: 10.1038/s41598-021-92718-7.
43. J. K. Kim, H. Y. Park, B. L. Lee, The symbiotic role of O-antigen of *Burkholderia* symbiont in association with host *Riptortus pedestris*. *Developmental & Comparative Immunology* 60, 202-208 (2016). doi: 10.1016/j.dci.2016.02.009.
44. B. Sit, V. Srisuknimit, E. Bueno, F. G. Zingl, K. Hullahalli, F. Cava, M. K. Waldor, Undecaprenyl phosphate translocases confer conditional microbial fitness. *Nature* 613, 721-728 (2023). doi: 10.1038/s41586-022-05569-1.
45. J. Szczepaniak, C. Press, C. Kleanthous, The multifarious roles of Tol-Pal in Gram-negative bacteria. *FEMS Microbiology Reviews* 44, 490-506 (2020). doi: 10.1093/femsre/fuaa018.
46. M. A. Valvano, Remodelling of the Gram-negative bacterial Kdo_2-lipid A and its functional implications. *Microbiology* 168, 001159 (2022). doi: 10.1099/mic.0.001159.
47. B. Obadia, Z. T. Güvener, V. Zhang, J. A. Ceja-Navarro, E. L. Brodie, W. W. Ja, W. B. Ludington, Probabilistic invasion underlies natural gut microbiome stability. *Current Biology* 27, 1999-2006 (2017). doi: 10.1016/j.cub.2017.05.034.
48. M. Zasloff, Antimicrobial peptides of multicellular organisms. *Nature* 415, 389-395 (2002). doi: 10.1038/415389a.
49. M. A. Hanson, L. Grollmus, B. Lemaitre, Ecology-relevant bacteria drive the evolution of host antimicrobial peptides in *Drosophila*. *Science* 381, eadg5725 (2023). doi: 10.1126/science.adg5725.
50. P. Mergaert, Y. Kikuchi, S. Shigenobu, E. C. M. Nowack, Metabolic integration of bacterial endosymbionts through antimicrobial peptides. *Trends Microbiology* 25, 703-712 (2017). doi: 10.1016/j.tim.2017.04.007.

Figures

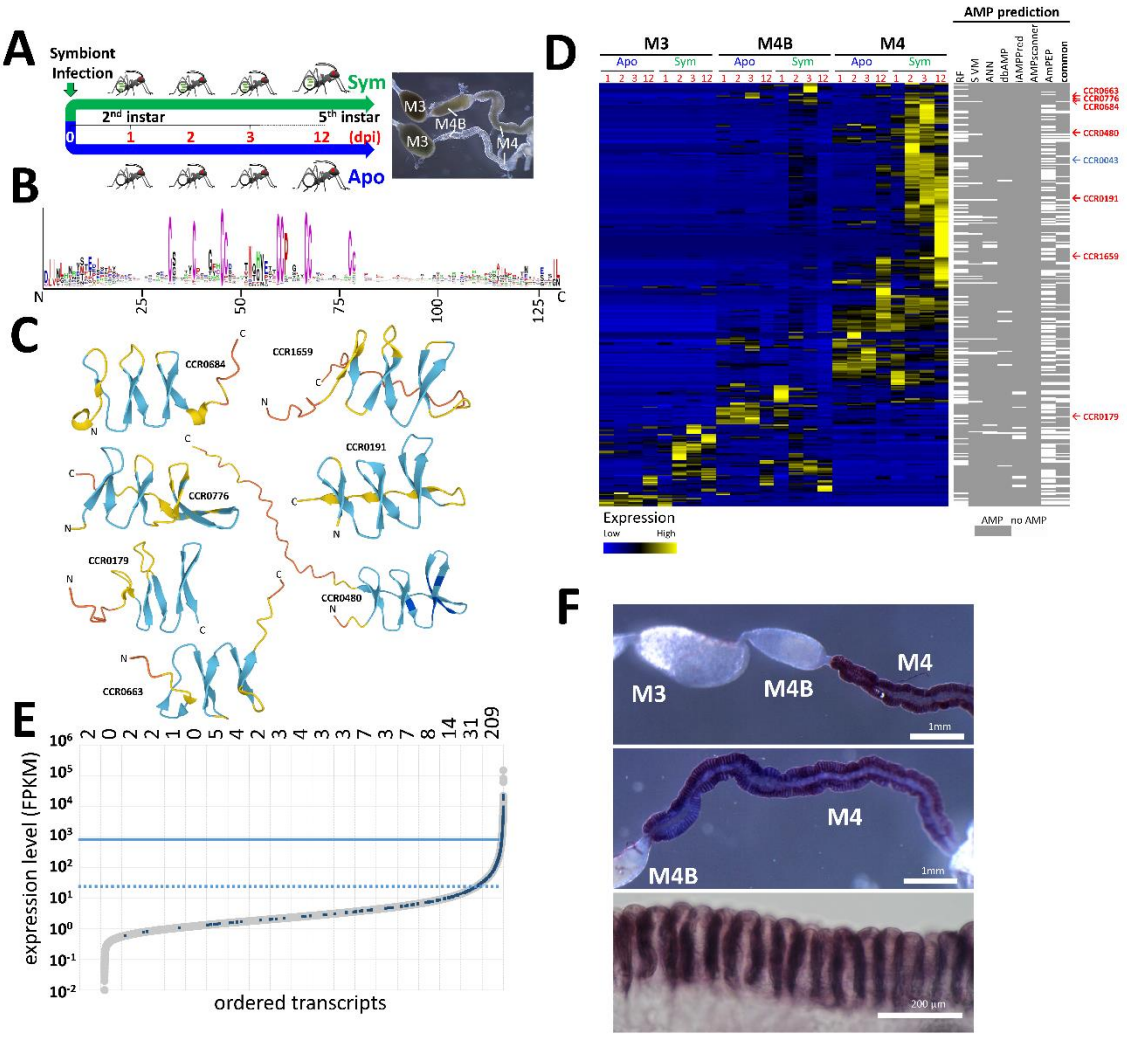
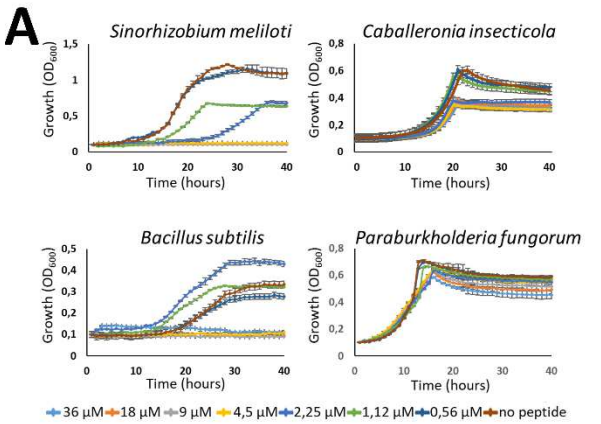


Figure 1. CCRs are symbiosis-specific AMP-like peptides. (**A)** Experimental setup for transcriptome analysis. First day second instars were divided in two groups. To one of them, *C. insecticola* symbionts were administered (green, Sym) and the other group remained free of symbionts (blue, Apo). Insects were dissected in the second (1, 2, and 3 days post inoculation [dpi]) or fifth instar (12 dpi) and the M3, M4B and M4 regions were harvested for transcriptome analysis. The pictures at the right show representative guts of a Sym insect at 3 dpi (top) and a same age Apo insect (bottom). (**B**) Logo profile of the mature CCR peptides identified in the transcriptome, highlighting the sequence diversity of the peptides and the ten conserved cysteine residues. (**C**) AlphaFold2 structural predictions of examples of CCR peptides showing antiparallel β-sheets carrying the cysteine residues. (**D**) Blue-black-yellow heat map of the relative expression profile of the identified CCR genes and white-grey heat map of AMP predictions. Sample identity in the expression heat map is indicated at the top and is according to panel A. AMP prediction tools are Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), AMPpredictor (dbAMP); iAMPpred; Antimicrobial Peptide Scanner (AMPscanner); (AmPEP_v1 and AmPEP_v2). A consensus prediction (6 out of 7 positive predictions) is indicated in the last column. The peptides used for functional characterization are indicated at the right of the heat maps. (**E**) Whole-mount *in situ* hybridization with a *CCR0043* antisense probe on the dissected

midgut of a 3 dpi symbiotic insect. Positive signal appears with a blue-brownish color. CCR0043 is specifically expressed in the M4 and uniformly in all crypts. Control *in situ* hybridizations on the gut of aposymbiotic insects and with a sense probe on symbiotic insects are shown in Supplemental Figure S2. (**F**) *CCR* transcript expression levels. Transcripts are ordered according to their expression level in the x-axis and their expression levels (FPKM, Fragments Per Kilobase of transcript per Million mapped reads) are plotted in the y-axis. All transcripts are indicated with grey dots and the *CCR* transcripts are indicated with blue crosses. The dotted and plain blue horizontal lines correspond to the mean expression level of all transcripts and *CCR* transcripts, respectively. The numbers above the plot indicate the number of *CCR* transcripts present in 5-percentile bins of transcripts. 77 % of the *CCR* transcripts are among the 10 % highest expressed transcripts in the midgut.



B

	Sinorhizobium meliloti	*Bacillus subtilis*	*Caballeronia insecticola*	*Paraburkholderia fungorum*
PMB	1,4	2,8	22,5	90
LL37	0,69	5,56	2,78	>5,56
NCR335	1,625	>13	3,25	>13
thanatin	1,875	120	>120	>120
riptocin	<0,78	>100	50	50
CCR1659	2,25	4,5	>36	>36
CCR0191	5	10	>40	>40
CCR0179	5	20	40	40
CCR0480	<0,84375	13,5	>27	>27
CCR0663	11,25	22,5	>45	>45
CCR0684	21	10,5	>42	>42
CCR0776	40	>40	>40	>40

resistant

sensitive

Figure 2. CCR peptides are AMPs. (**A**) Growth inhibition of the indicated bacterial species by different concentrations of CCR1659. Error bars are standard deviations (n=3). (**B**) Minimal concentrations (in µM) of growth inhibition of the indicated strains by various peptides.

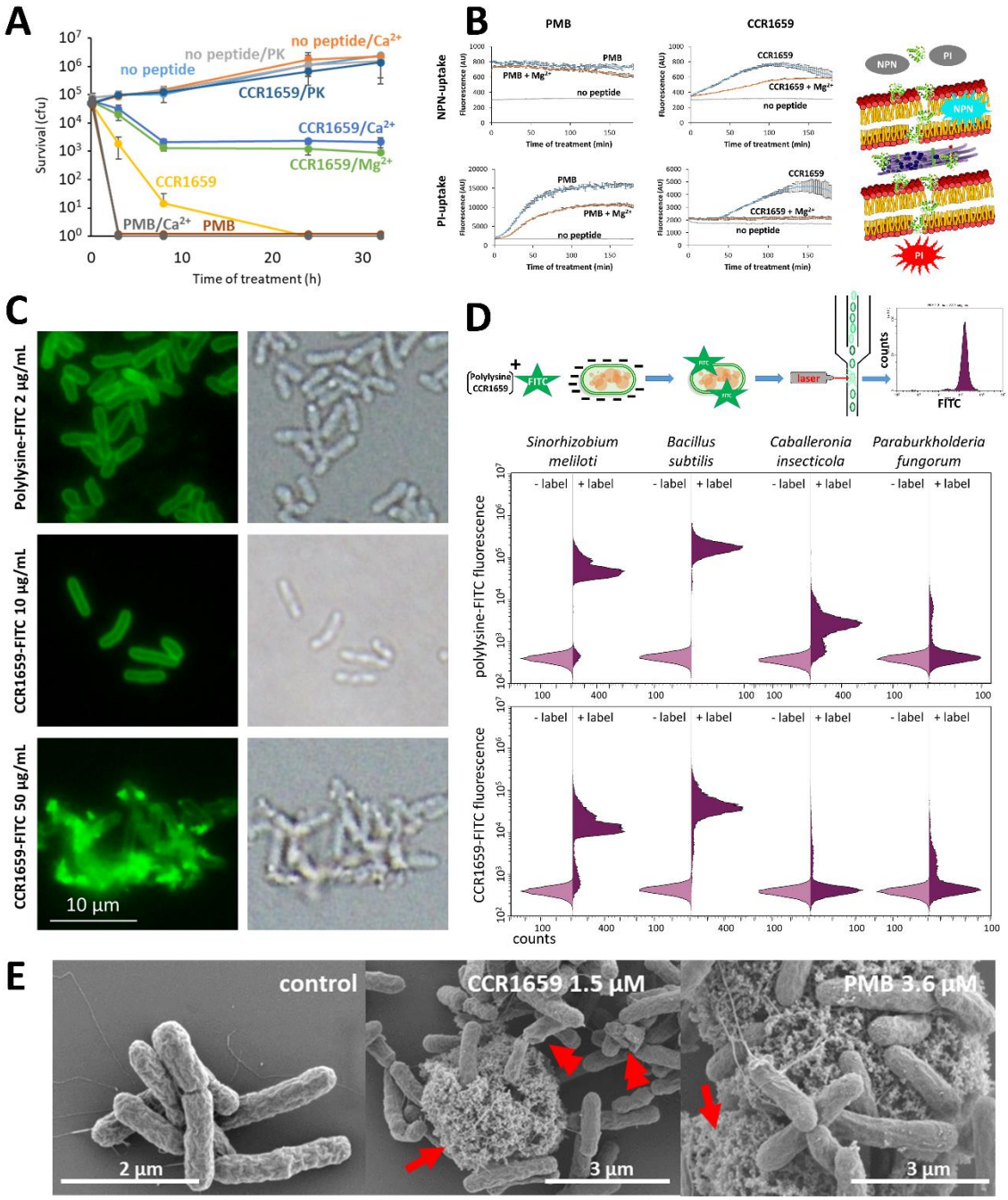


Figure 3. CCR peptides target membranes. (**A**) Bactericidal activity of 25 µM CCR1659 and 25 µM PMB. PK: proteinase K; Ca^{2+}: activity in the presence of 5 mM $CaCl_2$; Mg^{2+} activity in the presence of 5 mM $MgCl_2$. Error bars are standard deviations (n=3). (**B**) NPN and PI uptake by *S. meliloti* cells in response to treatment with 10 µM CCR1659 or 10 µM PMB in the presence or absence of 5 mM $MgCl_2$. NPN is a lipophilic dye that fluoresces in hydrophobic environments such as bacterial phospholipids exposed by outer membrane damage; PI is a membrane impermeant DNA-intercalating dye that fluoresces upon DNA binding in the cytoplasm, indicative of permeabilisation of both the outer and inner membrane. Error bars are standard deviations (n=3).

(**C**) Fluorescence microscopy (left) of *S. meliloti* cells treated with Polylysine-FITC or CCR1659-FITC at the indicated concentrations. Corresponding bright field images are shown at the right. (**D**) Flow cytometry analysis of Polylysine-FITC (top) or CCR1659-FITC binding by the indicated strains. Light purple histograms are control measurements without fluorescent label (-label); the dark purple histograms are in the presence of the fluorescent label (+label). (**E**) SEM micrographs of untreated *S. meliloti* cells (left) or treated with 1.5 μM CCR1659 (middle) or with 3.6 μM PMB (right). The arrows indicate cellular material released from cells. The double arrowheads indicate cells with lost turgor.

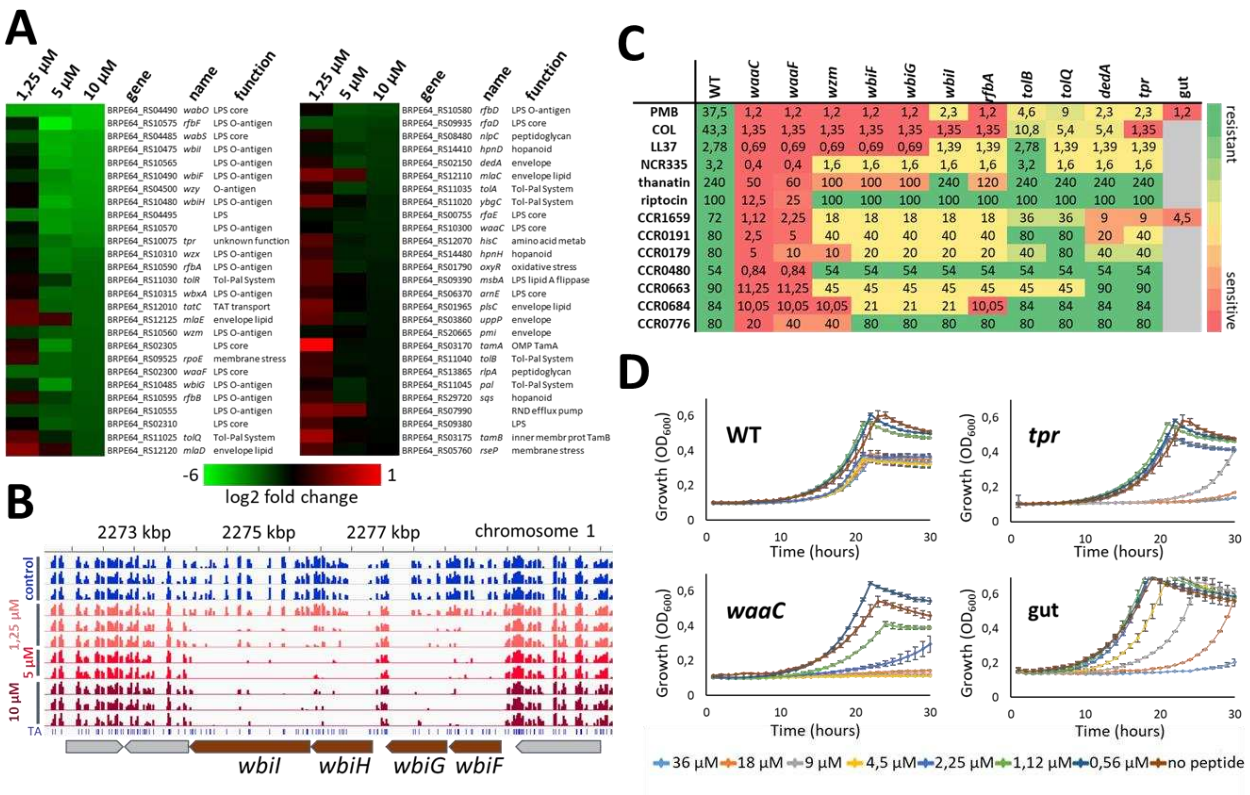
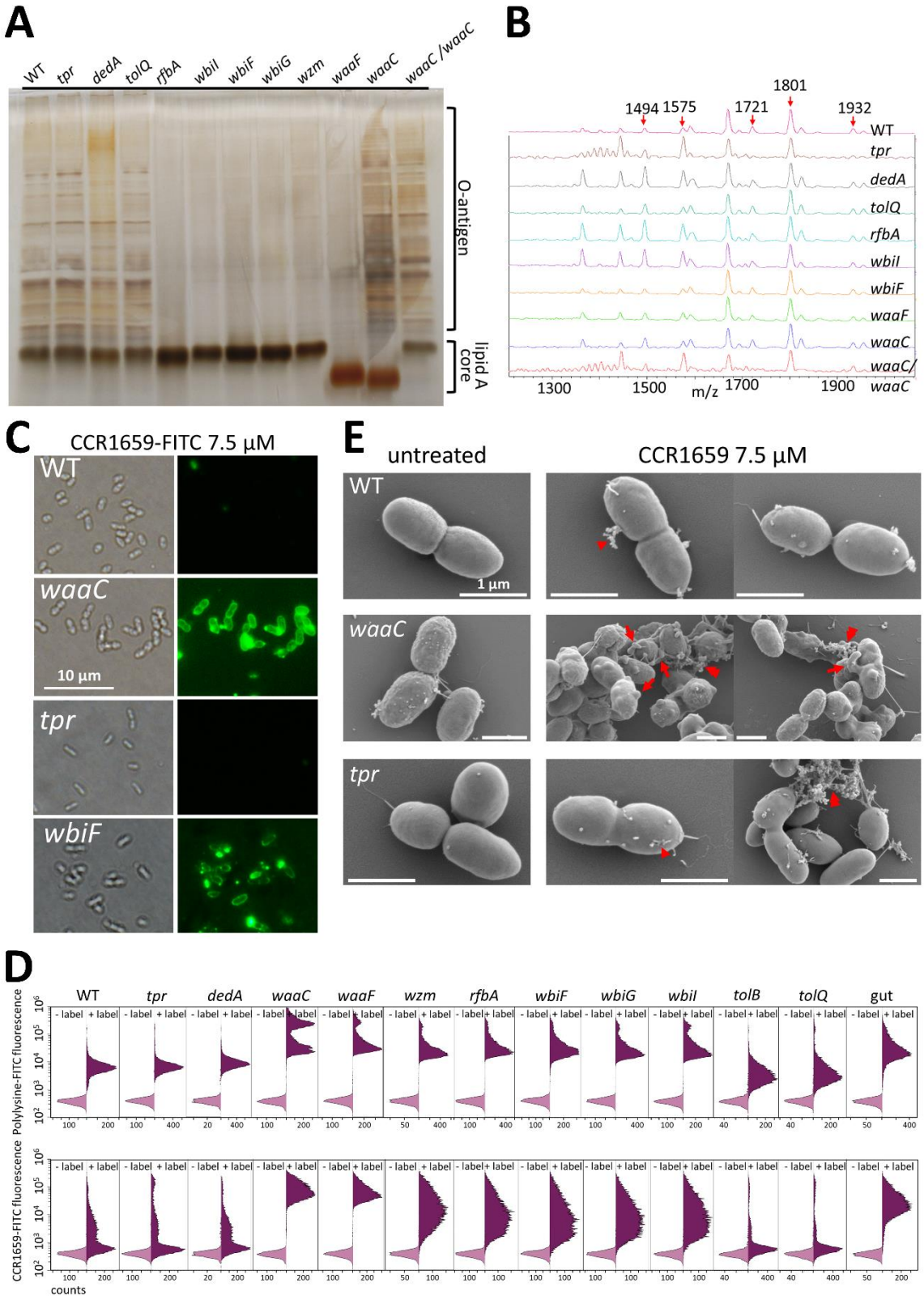


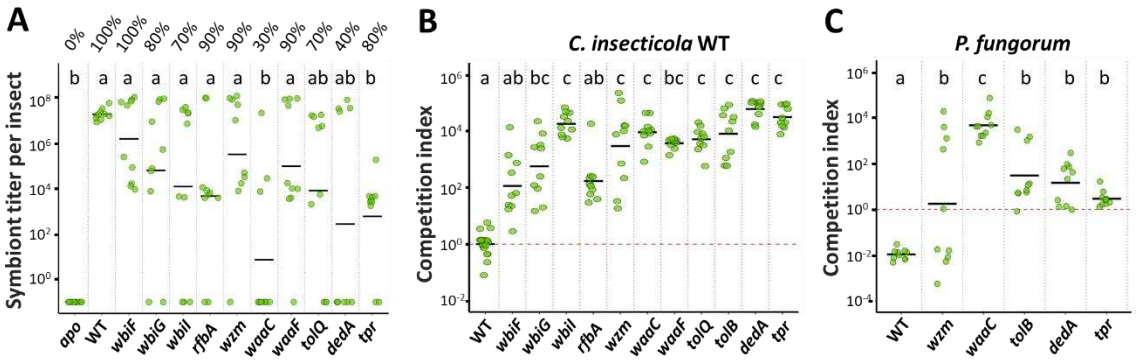
Figure 4. Identification of AMP resistance genes by Tn-seq. (**A**) Heat map showing the level of depletion of transposon insertions in the indicated genes in the *C. insecticola* population grown in the presence of PMB at the indicated concentrations. The color-code scale indicates the log2 fold change in the insertion abundance under the test conditions relative to the control conditions. (**B**) IGV view of Tn-seq sequencing data for a selected genomic region of *C. insecticola*. The histograms indicate the abundance of mutants in the Tn-seq population for the indicated samples. Genes whose products contribute to PMB resistance have a lower frequency of Tn insertions in peptide treatment screens than in the control. (**C**) Mutants in selected genes are hypersensitive to AMPs. Heat map and minimal concentrations of growth inhibition of the indicated wild-type and mutant strains by the listed peptides. Minimal concentrations are indicated in µM. The color key of the heat map is as indicated at the right. Grey cells indicate not tested. (**D**) Growth inhibition of the indicated strains by different concentrations of CCR1659. "Gut" in panels C and D indicates crypt-colonizing *C. insecticola* bacteria, directly isolated from dissected M4. Error bars are standard deviations (n=3).


A
WT
tpr
dedA
tolQ
rfbA
wbiI
wbiF
wbiG
wzm
waaF
waaC
waaC /waaC
O-antigen
lipid A core
B
1494
1575
1721
1801
1932
WT
tpr
dedA
tolQ
rfbA
wbiI
wbiF
waaF
waaC
waaC/ waaC
1300
1500
m/z
1700
1900
C
CCR1659-FITC 7.5 µM
WT
waaC
10 µm
tpr
wbiF
E
untreated
CCR1659 7.5 µM
WT
1 µm
waaC
tpr
D
WT
tpr
dedA
waaC
waaF
wzm
rfbA
wbiF
wbiG
wbiI
tolB
tolQ
gut
- label
+ label
Polylysine-FITC fluorescence
CCR1659-FITC fluorescence
counts

668

669 **Figure 5. Surface properties of the AMP-sensitive mutants.** (**A**) Polyacrylamide gel
670 electrophoresis analysis of total LPS extracted from the indicated strains. The *waaC*/*waaC* strain
671 is the complemented mutant. Despite the altered core in the *waaC* mutant, an O-antigen ladder is
672 visible, that has a similar profile to the wild type, possibly corresponding to the O-antigen anchored
673 on an intermediate lipid carrier. (**B**) MS analysis of the lipid A molecule present in the indicated
674 mutants. Red arrows indicate the Ara4N carrying lipid A (SI Appendix, Fig. S7). (**C**) Fluorescence
675 microscopy of *C. insecticola* wild-type, *waaC*, *tpr* and *wbiF* cells treated with 50 µg/mL CCR1659-
676 FITC. All images are at the same magnification and the scalebar is 10 µm. (**D**) Flow cytometry
677 analysis of 50 µg/mL Polylysine-FITC (top) or 7.5 µM CCR1659-FITC binding by the indicated
678 strains. "Gut" is bacteria directly isolated from the midgut crypts. Light purple histograms are control
679 measurements without fluorescent label (-label); the dark purple histograms are in the presence of
680 the fluorescent label. Note the presence of a double peak in the Polylysine-FITC treated mutants
681 *waaC*, *waaF*, *wzm*, *rfbA*, *wbiFGI*, indicating of a heterogeneous bacterial population. (**E**) SEM
682 micrographs of untreated *C. insecticola* wild type and *waaC* and *tpr* mutant untreated cells or
683 treated with 7.5 µM CCR1659. Arrowheads indicate release of tiny amounts of cellular material in
684 intact cells. Double arrowheads indicate cellular material released from lysed cells. Arrows indicate
685 cell deformations. Scale bars are 1 µm for all images.

686



687
688 **Figure 6. AMP-sensitive mutants are impaired in gut colonization. (A**) Single-strain infections
689 of *R. pedestris* second instar nymphs with *C. insecticola* WT or indicated mutants or no bacteria
690 (*apo*). Colonization of the M4 crypt region was determined at 5 dpi by dissection and microscopy
691 observation of the guts and symbiont titer determination by qPCR in M4 total DNA extracts. The %
692 above the dot plots indicate the proportion of insects that showed colonization by microscopy
693 observation (n=10). The qPCR measurements for each individual insect are indicated by green
694 dots and the mean per mutant is indicated by a horizontal black line. (**B**) Co-infections of *R.*
695 *pedestris* with an equal mix of RFP-labelled *C. insecticola* WT and indicated GFP-labelled WT or
696 mutant strains. Relative abundance of the two strains in the M4 midgut regions at 5 dpi was
697 determined by flow cytometry on dissected intestines. The competition index expresses for all
698 samples the ratio of RFP-labelled WT to the GFP-labelled WT or indicated mutant in the M4 migut
699 region, corrected by the ratio of the inoculum, which was in all cases close to 1. Each dot represents
700 the competition index in an individual and the mean per mutant is indicated by a horizontal black
701 line (n=10). (**C**) Co-infections of *R. pedestris* with a 1:1 mix of GFP-labelled *P. fungorum* and
702 indicated mScarlett-I-labelled WT or mutant *C. insecticola*. Relative abundance of the two strains
703 in the M4 midgut regions at 5 dpi was determined by flow cytometry on dissected intestines. The
704 competition index expresses for all samples the ratio of *P. fungorum* to the indicated mutant,
705 corrected by the ratio of the inoculum, which was in all cases close to 1. Each dot represents the
706 competition index in an individual and the mean per mutant is indicated by a horizontal black line
707 (n=10). In all panels, different letters indicate statistically significant differences ($P<0.05$). Statistical
708 significance was analyzed by Kruskal–Wallis test, Dunn *post hoc* test and Benjamini-Hochberg
709 correction.

Microbiology Resource Announcements

| Environmental Microbiology | Announcement

Complete and circularized genome sequences of five nitrogen-fixing *Bradyrhizobium* sp. strains isolated from root nodules of peanut, *Arachis hypogaea*, cultivated in Tunisia

Besma Bouznif,[1,2,3,4] Amira Boukherissa,[1,2] Yan Jaszczyszyn,[1] Mohamed Mars,[3] Tatiana Timchenko,[1] Jacqui A. Shykoff,[2] Benoît Alunni[1,5]

AUTHOR AFFILIATIONS See affiliation list on p. 3.

ABSTRACT This manuscript reports the complete and circularized Oxford Nanopore Technologies (ONT) long read-based genome sequences of five nitrogen-fixing symbionts belonging to the genus *Bradyrhizobium*, isolated from root nodules of peanut (*Arachis hypogaea*) grown on soil samples collected from Tunisia.

KEYWORDS symbiosis, nitrogen fixation, *Bradyrhizobium*, peanut, Tunisia

Legume plants have evolved the capacity to harbor symbiotic nitrogen-fixing soil bacteria within root nodules. This symbiosis provides these plants the ability to grow in diverse habitats and to become major players in agricultural sustainability (1, 2). Bacteria belonging to the genus *Bradyrhizobium* are known as predominant peanut (*Arachis hypogaea*) symbionts (3–6).

In this announcement, we report the complete genome sequence of five bacterial strains isolated from peanut root nodules and belonging to the genus *Bradyrhizobium* based on average nucleotide identity (ANI) value analysis (Table 1). The soils were sampled from five sites in Tunisia (Table 1). Then, Tunisian traditional varieties of peanuts were grown in pots containing the sampled soils for bacterial trapping. After 7 weeks of greenhouse cultivation (28°C, 16 h photoperiod, 160 µmol·m^{-2}·S^{-1}), nodules were collected and surface-sterilized in 96% ethanol for 1 min, then in 3% sodium hypochlorite solution for 3 min before being rinsed three times with sterile distilled water and individually crushed, plated on yeast extract mannitol (YM) agar medium and grown for 5 to 10 days at 28°C (7, 8). The strains were purified by successive streaking and single-colony picking and then stored in YM medium with 20% (vol/vol) glycerol at −80°C.

Total genomic DNA was extracted from the five strains grown in the YM medium for 4 days at 28°C using the MasterPure complete DNA and RNA Purification Kit (Epicentre). DNA libraries were prepared from intact genomic DNA with the Native Barcoding Kit 24 V14 (SQK-NBD114.24), and sequencing was performed on a MinION flow cell (R10.4.1). Basecalling was performed with the Super Accuracy model in Guppy 5.3.1. Nanopore library adaptors were trimmed using Porechop v2.0.4 (https://github.com/rrwick/Porechop). The read quality was assessed using NanoPlot v1.41.0 (9). The long reads were then assembled using Flye 2.8.3 (10), which generated a single contig assembly for four strains and two contigs for the strain BWA-3-5 (https://doi.org/10.6084/m9.figshare.24547348). The presence of plasmids was determined with plASgraph2 v1.0.0 (11). The assemblies were polished with Medaka v1.7.2 (model r1041_e82_260bps_sup_g632) (https://github.com/nanoporetech/medaka) using the Oxfod Nanopore Technology (ONT) reads to create a consensus sequence. The assemblies of the chromosomes were then rotated with Circlator all at the DnaA gene (12).

Editor Leighton Pritchard, University of Strathclyde, Glasgow, United Kingdom

Address correspondence to Benoît Alunni, benoit.alunni@inrae.fr.

The authors declare no conflict of interest.

See the funding table on p. 3.

Received 13 November 2023
Accepted 15 April 2024
Published 15 May 2024

TABLE 1 Genome features of the five *Bradyrhizobium* sp. strains

Strain characteristic(s)	Data for:				
	***Bradyrhizobium* sp. BEA-2-5**	***Bradyrhizobium* sp. BWA-3-5**	***Bradyrhizobium* sp. BWC-3-1**	***Bradyrhizobium* sp. NDS-1**	***Bradyrhizobium* sp. sBnM-33**
Sampling site	Béja, Elmarja	Béja, Wechtata	Béjà, Wechtata	Nabeul, Dar Allouch	Ben Gardane
Peanut variety	Arbi	Arbi	Chounfakhi	Siniya	Massriya
No. of reads (post QC)	91,274	118,365	141,671	99,886	80,650
No. of contigs	1	2	1	1	1
No. of genes	7,837	7,729	8,367	7,134	8,642
Genome length (bp)	8,406,336	8,041,125 (chromosome: 7,897,608; plasmid: 143,517)	8,767,305	7,645,890	9,184,954
Coverage (×)	148	188	161	190	146
N50 (bp)	8,406,336	7,897,608	8,767,305	7,645,890	9,184,954
GC content (%)	63.92	62.51	62.92	64.01	61.62
BUSCO score (%)	99.1	99.9	99.5	99.4	99.2
Closest taxonomic assignation—accession (ANI value to closest species [%])	*Bradyrhizobium pachyrhizi*—GCF_029714545.1 (95.89)	*Bradyrhizobium hereditatis*—GCF_020329435.1 (89.14)	*Bradyrhizobium canariense*—GCF_019402665.1 (94.70)	*Bradyrhizobium frederickii*—GCF_004570865.1 (90.13)	*Bradyrhizobium retamae*—GCF_001440415.1 (92.53)
BioSample ID	SAMN37684786	SAMN37684780	SAMN37683990	SAMN37684784	SAMN37684830
SRA ID	SRR26337671	SRR26335727	SRR26335726	SRR26337672	SRR26336054
GenBank ID	CP136629	CP136626-CP136627	CP136625	CP136628	CP136624

The plasmid of the strain BWA-3-5 was rotated with Prodigal v2.6.2 at position 73241 (13). The statistics of the assembly were obtained with QUAST v5.2.0 (https://github.com/ablab/quast), and the completeness of the genomes was assessed using BUSCO v5.6.1 (14) (Table 1). Functional annotation of the genomes was performed using the NCBI

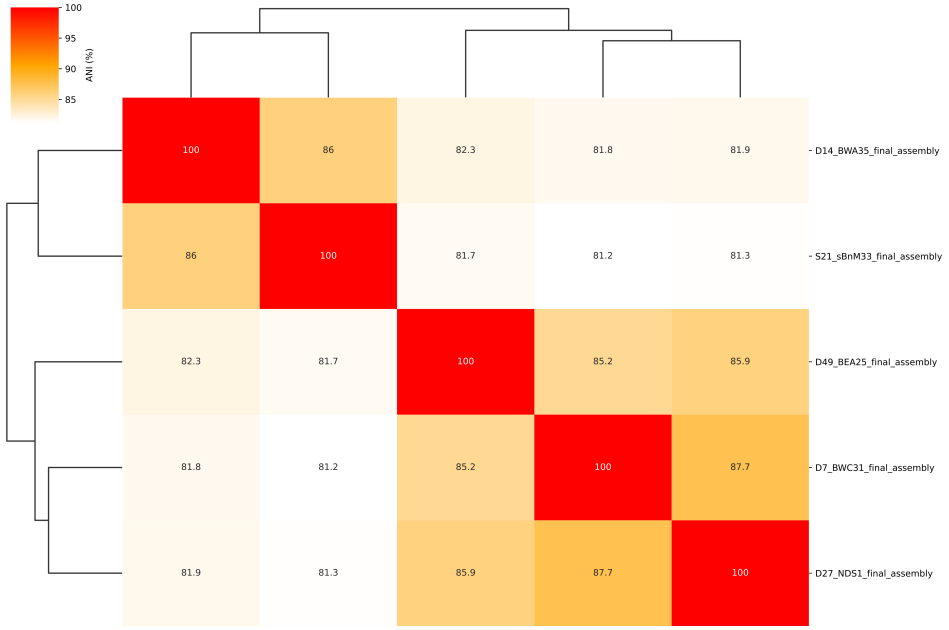


FIG 1 Heatmap of ANI values between the five *Bradyrhizobium* sp. strains. ANI values for pairwise comparisons of these five strains are in the range between 81.2% and 87.7%, below the species threshold of 95%, and therefore appear to belong to five distinct *Bradyrhizobium* species.

Prokaryotic Genome Annotation Pipeline v6.3 (15, 16). ANI calculation was carried out using ANIclustermap v1.2.o (Fig. 1)(17).

ACKNOWLEDGMENTS

B.B. benefited from support from the Programme Hubert Curien UTIQUE Project 17G0918, the Government of Tunisia, for a mobility grant (Bourse d'Alternance) and a doctoral partner scholarship from the French Ministry of Foreign and European Affairs. A.B. benefited from a Ph.D. contract in the frame of the CNRS 80|PRIME—2021 program. We acknowledge the sequencing and bioinformatics expertise of the I2BC High-throughput sequencing facility, supported by France Génomique (funded by the French National Program "Investissement d'Avenir" ANR-10-INBS-09). This work has benefited from a French State grant (Saclay Plant Sciences, reference no. ANR-17-EUR-0007, EUR SPS-GSR) under a France 2030 program (reference no. ANR-11-IDEX-0003).

AUTHOR AFFILIATIONS

[1]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France
[2]IDEEV—Laboratoire Ecologie, Systématique et Evolution, CNRS, AgroParisTech Université Paris-Saclay, Gif-sur-Yvette, France
[3]Research Unit Biodiversity and Valorization of Arid Areas Bioresources (BVBAA), Faculty of Sciences, Gabès, Tunisia
[4]Université de Picardie Jules Verne, UMRt BioEcoAgro 1158-INRAE, BIOPI, Amiens, France
[5]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), Versailles, France

AUTHOR ORCIDs

Benoît Alunni http://orcid.org/0000-0002-3616-0669

FUNDING

Funder	Grant(s)	Author(s)
Programme Hubert Curien	UTIQUE Project 17G0918	Bouznif Besma Mars Mohamed Shykoff Jacqui
Centre National de la Recherche Scientifique (CNRS)	80 Prime	Boukherissa Amira Benoît Alunni
Agence Nationale de la Recherche (ANR)	ANR-10-INBS-09	Jaszczyszyn Yan Benoît Alunni
Agence Nationale de la Recherche (ANR)	ANR-17-EUR-0007	Timchenko Tatiana Benoît Alunni
Agence Nationale de la Recherche (ANR)	ANR-11-IDEX-0003	Timchenko Tatiana Benoît Alunni

DATA AVAILABILITY

The raw reads and complete genomes of the five *Bradyrhizobium* sp. strains were deposited at GenBank under the BioProject no. PRJNA1023532. The BioSample numbers, raw reads, and assembly GenBank accession numbers are given in Table 1.

REFERENCES

1. Sprent JI. 2007. Evolving ideas of legume evolution and diversity: a taxonomic perspective on the occurrence of nodulation. New Phytol 174:11–25. https://doi.org/10.1111/j.1469-8137.2007.02015.x
2. Liu A, Ku Y-S, Contador CA, Lam H-M. 2020. The impacts of domestication and agricultural practices on legume nutrient acquisition through symbiosis with rhizobia and arbuscular mycorrhizal fungi. Front Genet 11:583954. https://doi.org/10.3389/fgene.2020.583954
3. Muñoz V, Ibañez F, Tonelli ML, Valetti L, Anzuay MS, Fabra A. 2011. Phenotypic and phylogenetic characterization of native peanut *Bradyrhizobium* isolates obtained from Córdoba, Argentina. Syst Appl Microbiol 34:446–452. https://doi.org/10.1016/j.syapm.2011.04.007
4. Jaiswal SK, Msimbira LA, Dakora FD. 2017. Phylogenetically diverse group of native bacterial symbionts isolated from root nodules of groundnut (*Arachis hypogaea L.*) in South Africa. Syst Appl Microbiol 40:215–226. https://doi.org/10.1016/j.syapm.2017.02.002
5. Shao S, Chen M, Liu W, Hu X, Wang E-T, Yu S, Li Y. 2020. Long-term monoculture reduces the symbiotic rhizobial biodiversity of peanut. Syst Appl Microbiol 43:126101. https://doi.org/10.1016/j.syapm.2020.126101
6. Bouznif B, Alunni B, Mars M, Shykoff JA, Timchenko T, Rodríguez de la Vega RC. 2021. Draft genome sequences of nitrogen-fixing *Bradyrhizobia* isolated from root nodules of peanut, *Arachis hypogaea*, cultivated in Southern Tunisia. Microbiol Resour Announc 10:e0043421. https://doi.org/10.1128/MRA.00434-21
7. Rodríguez-Echeverría S, Pérez-Fernández MA, Vlaar S, Finan TM. 2003. Analysis of the legume–rhizobia symbiosis in shrubs from central Western Spain. J Appl Microbiol 95:1367–1374. https://doi.org/10.1046/j.1365-2672.2003.02118.x
8. Vincent JM. 1970. A manual for the practical study of the root-nodule bacteria. Man Pract Study Root-Nodule Bact.
9. De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39:btad311. https://doi.org/10.1093/bioinformatics/btad311
10. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8
11. Sielemann J, Sielemann K, Brejová B, Vinař T, Chauve C. 2023. PlASgraph2: using graph neural networks to detect plasmid contigs from an assembly graph. Front Microbiol 14:1267695. https://doi.org/10.3389/fmicb.2023.1267695
12. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16:294. https://doi.org/10.1186/s13059-015-0849-0
13. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
14. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351
15. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105
16. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569
17. Lisa M, Bot H. 2017. My research software (2.0.4)

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences du végétal: du gène à l'écosystème (SEVE)

Titre : Étude de la coévolution entre peptides antimicrobiens et transporteurs de peptides dans le cadre la symbiose rhizobium-légumineuses

Mots clés : symbiose, phylogénie moléculaire, évolution, transporteur ABC, peptides antimicrobiens

Résumé :

Les légumineuses présentant une carence en azote peuvent entrer en interaction symbiotique avec des bactéries du sol fixatrices de N2 appelées rhizobia. Dans cinq clades de légumineuses, une stratégie d'exploitation appelée différenciation terminale des bactéroïdes (TBD) a évolué dans laquelle les rhizobiums subissent une différenciation extrême. Les bactéries terminalement différenciées sont plus grandes, polyploïdes, ont une membrane perméabilisée, et sont meilleures à la fixation de N2, fournissant un retour sur investissement plus élevé pour la plante. Nous savons que dans deux clades, IRLC (par exemple, Medicago spp.) et Dalbergioïds (par exemple, Aeschynomene spp.), ce processus de différenciation est déclenché par un ensemble de peptides antimicrobiens végétaux apparemment non apparentés avec une activité antimicrobienne à la membrane connue sous le nom de peptides Nodule-spécifiques Cystéine-Riche (NCR). À son tour, les rhizobia exposés au stress provoqué par les NCRs nécessitent un transporteur de peptides ABC de la famille BacA pour faire face à ce stress. Cependant, si des peptides NCR ou des peptides similaires sont également trouvés dans d'autres clades où la TBD se produit et la relation évolutive entre ces peptides reste inconnue. Dans ce projet, nous avons testé l'hypothèse d'une coévolution convergente entre les différents clades de légumineuses et leur rhizobia engagés dans ce programme de différenciation, tant au niveau phénotypique que moléculaire. Pour ce faire, nous avons combiné des analyses d'évolution moléculaire avec des tests fonctionnels, fournissant ainsi des connaissances expérimentales sur la question fondamentale de la contingence et de répétabilité en évolution tout en générant simultanément de nouveaux outils pour concevoir une symbiose plus efficace.

Title: Study of the coevolution between antimicrobial peptides and peptide transporters in legume-rhizobium symbiosis

Keywords: molecular phylogeny, symbiosis, ABC transporter, evolution, antimicrobial peptides

Abstract:

Legume plants under nitrogen deficiency can enter a symbiotic interaction with N_2-fixing soil bacteria called rhizobia. In five legume clades, an exploitive strategy called Terminal Bacteroid Differentiation (TBD) has evolved in which rhizobia undergo extreme differentiation. Terminally differentiated bacteria are larger, polyploid, have a permeabilized membrane, and are better at N_2 fixation, providing a higher return on investment for the plant. We know that in several members of the distantly related Inverted Repeat Lacking Clade (IRLC, e.g., *Medicago* spp.) and the Dalbergioid clade (e.g., *Aeschynomene* spp.), this differentiation process is triggered by a set of apparently unrelated plant antimicrobial peptides with membrane damaging activity known as Nodule-specific Cysteine-Rich (NCR) peptides. In turn, rhizobia exposed to NCR stress requires an ABC peptide transporter of the BacA family to cope with this stress. However, whether NCR peptides or similar peptides are also found in other clades where this occurs and the evolutionary relation among these peptides remain unknown. In this project, we tested whether NCR peptides and BacA peptide transporters evolved independently in the different legume clades that induce TBD and their rhizobia, implying convergent coevolution, both at phenotypic and molecular levels. We combined molecular evolution analyses with functional assays, thus providing experimentally informed knowledge on the fundamental question of the part of contingency and repeatability in evolution while simultaneously generating new tools to engineer a more efficient symbiosis.