Étude des transferts horizontaux de gènes chez les nématodes phytoparasites par l'exploitation de métagénomes du sol

Carole Belliardo

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE DE DOCTORAT

Etude des transferts horizontaux de gènes chez les nématodes phytoparasites par l'exploitation de métagénomes du sol

---

Carole Belliardo

Ecole doctorale ED 85 : Science de la vie et de la santé

Unité de recherche : Institut Sophia-Agrobiotech,
UMR INRAE 1355, CNRS 7254

Équipe « Genomics & Adaptive Molecular Evolution »

Présentée en vue de l'obtention
du grade de docteur en Biologie des interactions et écologie d'Université Côte d'Azur

Dirigée par : Dr. Etienne G.J. DANCHIN

Co-encadrée par : Dr. Marc BAILLY-BECHET

Co-encadrée par : Dr. Mathilde CLÉMENT

Soutenue le : 16 décembre 2022

Devant le jury, composé de :

Dr. Karine VAN DONINCK | Présidente

Dr. Marie-Nöelle ROSSO | Rapportrice

Dr. Christophe MOUGEL | Rapporteur

Dr. Laura EME | Examinatrice

Dr. Samuel MONDY | Examinateur

Dr. Justine LIPUMA | Membre invité

INRAE
la science pour la vie, l'humain, la terre

MYCOPHYTO
Cultivons l'équilibre de la nature

Etude des transferts horizontaux de gènes chez les nématodes phytoparasites par l'exploitation de métagénomes du sol

Jury :

Présidente du jury

Dr. Karine VAN DONINCK, Professeure, Université libre de Bruxelles (Bruxelles, Belgique)

Rapporteurs

Dr. Marie-Nöelle ROSSO, DR INRAE, Directrice d'unité BBF INRAE - Aix-Marseille Universités (Marseille, France)

Dr. Christophe MOUGEL, DR INRAE, IGEPP (Rennes, France)

Examinateurs

Dr. Laura EME, CR CNRS, Université Paris Saclay (Paris, France)

Dr. Samuel MONDY, Ingénieur de recherche Responsable de la plateforme Genosol, INRAE Bourgogne-Franche-Comté (Dijon, France)

Co-encadrants de thèse

Dr. Mathilde CLÉMENT, Responsable R&D, MYCOPHYTO, (Biot, France)

Dr. Marc BAILLY-BECHET, Maître de conférences, Université Côte d'Azur (Nice, France)

Directeur de thèse

Dr. Étienne DANCHIN, DR INRAE, Institut Sophia Agrobiotech (Biot, France)

Résumé

Les nématodes phytoparasites (NPP) sont parmi les plus importants ravageurs des cultures et menacent l'approvisionnement alimentaire mondial. Outre la nécessité de comprendre la biologie de ces organismes pour développer de nouvelles stratégies de lutte, ces nématodes sont fascinants en termes d'évolution génomique. Le parasitisme des plantes a évolué plusieurs fois indépendamment chez les nématodes selon des processus évolutifs convergents. Il semble que tous les NPP aient acquis des gènes bactériens et fongiques par transferts horizontaux de gènes (THG). Certains des gènes acquis horizontalement sont impliqués dans des fonctions parasitaires essentielles comme la dégradation des parois cellulaires des plantes ou l'assimilation des nutriments provenant des plantes. Cependant, plusieurs questions majeures restent encore en suspens concernant l'origine de ces gènes, leur distribution dans les génomes et la chronologie des événements d'acquisition. La plupart des NPP vivent dans le sol; nous pouvons donc supposer que ces gènes proviennent des micro-organismes telluriques. Cependant, la sous-représentation de ces micro-organismes dans les librairies de séquences généralistes a probablement limité les précédentes analyses sur les THG. Pour pallier ce problème, nous avons constitué une bibliothèque de protéines provenant de plus de 6 800 métagénomes du sol disponibles publiquement. Un problème important dans les métagénomes concerne la qualité des données provenant des organismes eucaryotes en raison de l'utilisation d'outils dédiés aux génomes procaryotes. Afin de mieux représenter le pool de gènes présents dans les environnements naturels des NPP, nous avons identifié les contigs eucaryotes et re-prédit les gènes et protéines en utilisant un prédicteur de gènes eucaryotes. Nous avons ainsi obtenu une librairie de protéines fiable et non redondante plus représentative de la biodiversité naturelle du sol.

En utilisant cette librairie enrichie en protéines de sol, nous avons effectué une détection de THG sur 18 génomes de NPP du clade *Tylenchina*. Ces organismes présentent des modes de parasitisme très diversifiés. Après curation manuelle, la proportion de gènes acquis par transferts horizontaux avec confirmation phylogénétique est comprise entre 0,5 et 1,9% des gènes codant pour des protéines. Les THG que nous avons détectés dans les génomes de NPP proviennent principalement de bactéries. Nous avons également observé des THG provenant d'organismes eucaryotes tels que des champignons et pour la première fois des protistes et des plantes. Les taxa les plus représentés parmi les donneurs sont les clades bactériens *Burkholderiaceae*, *Proteobacteria*, *Actinobacteria*, *Rhizobiales* et fongiques *Dikarya* qui comprennent de nombreuses espèces vivant dans le sol. L'utilisation de données métagénomiques a permis de préciser l'histoire des THG déjà décrits mais aussi d'identifier des centaines de nouveaux THG. Les prédictions fonctionnelles des THG nouvellement identifiées indiquent une large diversité de fonctions potentielles dont les implications biologiques pourront être plus précisément décrites dans le cadre d'expériences biochimiques. L'intégration de données environnementales dans notre librairie de référence a permis d'étendre la détection des THG et de compléter le catalogue des descendants des potentiels donneurs.

Mots clefs : métagénomique, transfert horizontal de gènes, micro-organismes, parasites, symbiotes, nématode.

Abstract

Plant-parasitic nematodes (PPN) are among the most important crop pests and threaten the world's food production. Besides the need to understand their biology to develop new control strategies, they are fascinating organisms in terms of genomic evolution. Plant parasitism has evolved several times independently in nematodes with some convergent evolutionary processes. For instance, all studied PPN have acquired bacterial and fungal genes by horizontal gene transfers (HGT). Some of the acquired genes are involved in essential parasitic functions like plant cell wall degradation or processing nutrients from the plant. However, several major questions concerning their origin, evolutionary fate and distribution in the genomes and timing of acquisition events remain unsolved. Most PPN live in soil; thus, we hypothesised that these genes originated from soil-dwelling microorganisms. However, the underrepresentation of soil microorganisms in generalist sequence libraries has previously limited HGT analyses.

To circumvent this problem, we built a protein library including more than 6,800 soil metagenomes from the Joint Genome Institute's IMG/M server. The first challenge was to make this massive dataset more accurate and suitable for HGT analysis in PPN genomes. An important issue in metagenomic data is the underrepresentation of eukaryotes and their annotation with prokaryotic tools. To better represent the pool of genes present in the natural environments of PPN, we identified eukaryotic contigs and re-predicted proteins using Augustus, a eukaryotic dedicated gene predictor. Moreover, we reduced the protein sequence redundancy and refined the taxonomic assignment. After all these steps, we obtained an improved and non-redundant database that was more representative of the soil's natural biodiversity. This soil protein library, two times larger than the classic library, contains mainly organisms genetically divergent than lab-cultured.

Then, we performed an HGT detection on proteins from 18 plant-parasitic nematode genomes of the *Tylenchina* clade, constituting a highly diverse group of PPN phenotypes, against our library enriched with soil protein. After manual curation, the proportion of protein-coding genes acquired by horizontal transfers with phylogenetic confirmation is between 0.5 to 1.9%. Those genes mainly originate from bacteria, but we also observed HGT from eukaryotic kingdoms such as fungi, protists and plants. The most represented taxa in donors are bacterial clades *Burkholderiaceae*, *Proteobacteria*, *Actinobacteria*, *Rhizobiales* and the fungal clade *Dikarya*. The usage of metagenomic data clarified the history of previously described HGTs but also identified hundreds of new HGTs. Functional analyses of the newly identified HGTs indicate a wide diversity of potential functions whose biological implications can be more precisely described in in-vitro experiments. Integrating environmental data in our reference library has allowed us to extend the detection of HGTs and to complete the catalog of potential donor offspring.

Mots-clés : metagenomics, horizontal gene transfers, microorganisms, parasites, symbionts, nematodes.

Remerciements

Ces trois années de thèse sont passées extrêmement vite et pourtant la liste des personnes à remercier est longue. Je suis extrêmement reconnaissante envers l'ensemble des personnes qui ont contribué à la bonne réalisation de ce projet, qui m'ont transmis les savoirs nécessaires pour accéder au grade de docteur mais aussi qui m'ont accompagnée et soutenue tout au long de ce processus.

En premier lieu, je souhaite remercier mes encadrants pour leur soutien sans faille, leur grande disponibilité. Je vous remercie de m'avoir guidée avec pertinence et délicatesse tout en me laissant la liberté de diriger ce projet.

Je remercie mon directeur de thèse, le Dr. Étienne Danchin, de m'avoir donné l'opportunité de découvrir la réalité du métier de chercheur en stage puis en thèse. C'est une grande chance et un réel plaisir d'être dirigée par une personne aussi passionnée, talentueuse et bienveillante. Tout au long de ce projet, tu as su m'aider à garder le cap et me donner, avec justesse, le tempo pour mener ce projet à bien dans le délai imparti. Tu es le directeur de recherche et le chef d'équipe idéal; tu inspires l'admiration et cela donne envie de te suivre.

Je remercie aussi le Dr Marc Bailly-Bechet qui a codirigé ce projet avec beaucoup de complémentarité. Cela a été très enrichissant de bénéficier de ta vision du projet et de tes conseils toujours originaux et exposés avec beaucoup de pédagogie. Tu es un enseignant-chercheur extraordinaire.

Je remercie aussi le Dr. Mathilde Clément d'avoir codirigé ce projet. Je te remercie de m'avoir permis de participer à divers projets menés par le pôle R&D. Collaborer avec toi est une expérience très stimulante et enrichissante que j'espère renouveler.

Je tiens aussi à remercier chaleureusement l'ensemble des membres du jury pour avoir accepté d'évaluer mes travaux de thèse et pour l'attention portée à mon travail.

Je remercie mes rapporteurs, le Dr. Marie-Nöelle Rosso et le Dr. Christophe Mougel pour la lecture attentive de mon manuscrit et le temps consacré à la rédaction de rapports détaillés.

Je remercie aussi mes examinateurs, le Dr. Laura Eme et le Dr. Samuel Mondy d'avoir participé avec mes rapporteurs à une discussion très riche et stimulante. Je vous suis aussi très reconnaissante d'avoir suivi mon travail tout au long de ce projet.

Et enfin, je remercie le Pr. Karine Van Doninck d'avoir accepté de présider mon jury et d'avoir participé activement à la discussion.

Je remercie la société *MYCOPHYTO* et le département *Santé des Plantes et Environnement* de l'*INRAe* qui ont cru en ce projet et qui l'ont soutenu durant ces trois années.

Je remercie tout particulièrement MYCOPHYTO et sa directrice générale, le Dr. Justine Lipuma, d'avoir sélectionné ma candidature pour ce poste de doctorat et de m'avoir ainsi permis de réaliser une thèse, CETTE thèse. Justine, tu es une personne très inspirante et je suis honorée d'avoir pu travailler avec toi.

Je tiens à remercier l'ensemble des membres de l'Institut Sophia Agrobiotech et plus particulièrement le Dr. Philippe Castagnone, Directeur d'institut Sophia Agrobiotech, de m'avoir accueillie dans son unité de recherche.

Je remercie le Dr. Pierre Abad et l'ensemble des membres de l'équipe Interactions Plantes-Nématodes pour leur accueil chaleureux. Je remercie particulièrement Caroline Caporalino et Nathalie Marteau de m'avoir amenée sur le terrain.

Je remercie très chaleureusement l'ensemble des membres du plateau de bioinformatique, Corinne Rancurel, Martine Da Rocha et Arthur Père, pour la mise en place et la maintenance des ressources qui m'ont permis de travailler dans des conditions optimales. Je remercie tout particulièrement Corinne pour sa collaboration sur ce projet, sa disponibilité et ses talents de développeuse.

Je remercie l'ensemble des membres de l'équipe *Génomique et Evolution Moléculaire Adaptative* et notamment Georgios, Djampa, Anna, Joffrey et Marine et les nombreux stagiaires dont j'ai eu la chance de partager le bureau. Vous m'avez soutenue avec de sages conseils, des pauses-café et des incroyables choux matcha***.

Je remercie le Dr Dominique Colinet pour la relecture de mon manuscrit de thèse et pour ses précieux conseils lors de nos échanges tout au long de ce projet. Je te remercie pour ta bienveillance et ton soutien.

Je remercie aussi le Dr. Karine Robbe pour la qualité de ses enseignements qui m'ont permis de découvrir la bioinformatique avec beaucoup de pédagogie et de clarté. Apprendre la programmation me semblait être un véritable challenge mais avec des cours d'une si grande qualité ce fut un grand plaisir.

Je remercie plus généralement l'ensemble du corps enseignant de l'université de Nice pour la qualité des cours dispensés. Pour finir, je remercie le Dr. Didier Forcioli de m'avoir rassurée et encouragée à poursuivre dans cette voie.

Je remercie le Dr. Clément Gilbert de m'avoir permis d'approfondir mes connaissances sur la question des transferts horizontaux et d'acquérir une première expérience de publication scientifique. Je te remercie pour ton soutien, tes encouragements et tes précieux conseils. C'est toujours un grand plaisir de collaborer avec un chercheur aussi passionné.

Je remercie le Dr Monique Revel Gnilka pour sa gentillesse et ses conseils.

Je remercie toute ma famille de m'avoir soutenue. Je remercie mes parents de m'avoir soutenue dans tous mes choix professionnels comme personnels. Je remercie particulièrement ma mère de m'avoir transmis sa passion pour les sciences et son "esprit logique". Je remercie mon frère, ma sœur et mon beau-frère pour leur soutien infini. J'exprime toute ma gratitude à Yohan qui a toujours été très patient et disponible. Je remercie aussi (par âge pour éviter les débats) Micka, Jess, John, Morgane et Romain. Je remercie aussi Keyla, Kara, Leyvin, Jayden, Kalvin, Luna, Eden, Kylian et Louisy; les moments passés avec vous sont toujours des bouffées d'air.

Je remercie plus largement toute ma famille, mes tantes et mes oncles. J'exprime toute ma gratitude envers Luce et Robert qui m'ont fait confiance.

Je remercie mes amis pour leur soutien et les longues heures où ils ont subi mes histoires de nématodes en croyant qu'il s'agissait de simples vers de terre alors que pas du tout!!! Je remercie particulièrement Karine et Xavier pour ces moments d'escapades ludiques (dont on est sortis victorieux!!!).

Je remercie Kévin, Loan, Hussam et Mickaël pour tous les moments que l'on a partagés durant ces années universitaires. Je n'étais pas là pour ça haha, mais je suis bien contente de vous avoir trouvé.

Je remercie Laetitia pour son soutien même de l'autre bout du globe, Anaïs pour sa passion pour les champignons et son enthousiasme pour mon sujet de thèse, Sara pour son amitié intemporelle et évidemment Anthony.

Pour finir, je remercie Julia. Heureusement, le commis a dépassé le chef mais le docteur a dépassé l'éternelle doctorante. Trève de plaisanterie, il n'y a pas de mots pour exprimer toute ma gratitude. Tu sais que je sais que tu sais…

Je remercie ma grand-mère qui m'a toujours encouragée jusqu'à son dernier souffle et mon père pour son amour inconditionnel.

Table des matières

Abréviations

CMA : Champignons mycorhiziens à arbuscules

ET : Éléments transposables

GH : Glycosides hydrolases

THG : Transferts horizontaux de gènes

NPP : Nématodes parasites de plantes

NG : Nématodes à galles

NK : Nématodes à kystes

Préambule

L'ensemble des ces travaux de recherche ont été co-financés par la société Mycophyto et le département santé des plantes et environnement de l'INRAE.

Au cours du XXIème siècle, le secteur agricole devra connaître de profonds changements pour faire face à l'augmentation démographique et aux problèmes environnementaux. La croissance démographique va se poursuivre durant les prochaines décennies, et la population mondiale pourrait atteindre 10 milliards d'humains entre 2050 et 2100.

Outre la nécessité de modifier les schémas économiques et sociaux de distribution et de consommation alimentaire, nous devons augmenter la production de 25 à 70% selon certains modèles pour garantir la sécurité alimentaire universelle. L'autre élément crucial, à prendre en considération dans le développement des futures méthodes agricoles, est l'impact environnemental. Les pratiques agricoles conventionnelles ont permis d'accroître les rendements mais cela s'est fait au détriment de l'environnement. En effet, ces méthodes dégradent la qualité de l'eau, des sols et de la biodiversité en détruisant les écosystèmes. Le secteur agricole a un impact important sur l'environnement et les changements climatiques, cependant ce secteur est le premier touché par ces changements.

Pour répondre aux exigences quantitatives et qualitatives de production, le secteur agricole doit optimiser le rendement et diminuer les pertes. Une grande partie des pertes agricoles sont dues aux ravageurs de cultures. Cependant, les

méthodes de lutte qui recourt à l'utilisation de pesticides toxiques pour l'environnement et la santé humaine doivent être proscrites. Les apports utilisés en matière de stimulation du développement végétal saturent les sols en azote et phosphore, ce qui dégrade la qualité biotique et abiotique des sols. Il est donc urgent de développer des pratiques agricoles spécifiques et durables.

On sait depuis longtemps que le sol joue un rôle clé sur le développement de la plante et dans son interaction avec l'environnement, même si la compréhension des processus biologiques en jeu reste limitée. Le sol représente un milieu de vie riche en éléments chimiques indispensables à la croissance végétale mais aussi en micro-organismes qui interagissent directement avec les plantes. Ce compartiment a longtemps été négligé et sa biodiversité était assez mal connue, notamment en raison de méthodes d'exploration limitées tel que sa composition en micro-organismes difficilement cultivables en laboratoire. Aujourd'hui, le développement technologique, et notamment le séquençage haut débit, a permis l'émergence d'approches métagénomiques permettant de mieux appréhender la richesse et la diversité des communautés microbiennes dans leur globalité.

La métagénomique nous offre un nouvel angle d'étude pour comprendre l'effet direct et indirect des micro-organismes présents naturellement dans les sols sur la santé des plantes. Cet effet peut être bénéfique dans le cas d'une relation symbiotique avec la plante comme les champignons mycorhiziens qui aide le développement racinaire et contribue à la santé des plantes. En revanche, de nombreux organismes et micro-organismes parasites ou pathogènes présents dans les sols ont un effet néfaste sur la plante, ce qui impacte directement les rendements. Il semble que les changements climatiques favorables à la prolifération de ravageurs de cultures risquent d'amplifier ces problèmes. C'est pourquoi il est urgent de mieux comprendre les processus biologiques impliqués dans cette interaction afin d'envisager des solutions ciblées.

Une grosse partie des pertes agricoles sont dues aux infections par les nématodes parasites de plantes qui détournent les nutriments des plantes et réduisent ainsi les rendements. Les produits phytosanitaires actuellement disponibles sont hautement toxiques pour la santé humaine et l'environnement. Cependant, sans ces produits, l'impact de ces ravageurs serait drastiquement amplifié. Pour développer de nouvelles méthodes de contrôle spécifiques et efficaces, nous avons besoin de mieux comprendre la biologie de ces organismes. On sait déjà que ces nématodes ont acquis des gènes impliqués dans le parasitisme par transferts horizontaux. De tels flux de gènes sont retrouvés plus largement chez d'autres phytopathogènes tels que les insectes ou les Oomycètes. Il semble que ce phénomène de transfert horizontal de gènes est impliqué dans l'émergence du parasitisme des plantes pour diverses lignées, et qu'il constitue ainsi un champ de recherche essentiel pour la compréhension du phyto-parasitisme.

Dans ce contexte, il semble essentiel de mieux comprendre l'écologie et l'évolution des organimes associés aux plantes. Durant cette thèse, deux axes d'étude ont été développés a partir de l'exploitation de données métagénomiques de sol. Un axe de recherche visait à étudier l'évolution des génomes des nématodes parasites de plantes, et plus particulièrement la manière dont les micro-organismes telluriques ont pu contribuer à l'adaptation d'un mode de vie parasitaire chez les nématodes via un flux de gènes. Ce premier axe s'inscrit dans les thématiques de recherche développées par le département Santé Des Plantes et Environnement de l'institut nationale de recherche pour l'agriculture, l'alimentation et environnement (INRAe) qui s'interesse aux questions relatives à la santé des cultures. D'autre part, un axe de recherche plus appliqué visait a extraire des connaissances concernant l'écologie des champignons mycorhyziens à arbuscules dans les sols afin d'alimenter les bases de données utilisées pour les projets de Recherche et développement de la société Mycophyto.

Introduction

1. Contexte général

A. Agriculture et enjeux actuels

1. Agriculture et impact démographique

La production agricole a un impact important sur le développement et la stabilité des sociétés humaines. L'émergence de l'agriculture durant la période du Néolithique constitue un moment clé de l'histoire humaine qui a bouleversé les modes de vie. Cette période a marqué la fin de la préhistoire avec une transition vers une économie de production reposant sur l'aménagement d'espaces de culture et la domestication des espèces. Au fil du temps, les agriculteurs ont développé des techniques, des outils et des machines pour optimiser les méthodes de production. Les données archéologiques mettent en évidence l'émergence simultanée et indépendante de l'agriculture dans au moins sept régions du globe. En Europe,

l’agriculture semble avoir émergé en Anatolie, une région correspondant à l'actuelle Turquie. (Marchi et al., 2022).

Le développement de l’agriculture a induit la sédentarisation des populations humaines qui se sont structurées localement et ont commencé à croître de plus en plus rapidement (Sahlins, 2011; Durand, 1979). À partir du XXème siècle, l'évolution de l’agriculture s’est accélérée grâce aux progrès réalisés notamment dans le domaine de la chimie permettant de développer les produits de synthèse à la base de l’agriculture conventionnelle.

2. *Agriculture et impact environnemental*

L'agriculture conventionnelle est un système de production caractérisé par l'usage d'intrants chimiques et de machinerie lourde permettant une production rapide et quantitative de denrées alimentaires. En revanche, de nombreux effets néfastes à la santé humaine et à l'environnement ont été recensés.

Les méthodes déployées en agriculture conventionnelle impactent profondément la qualité des sols et de l’environnement car les intrants s’accumulent et dégradent les milieux naturels. Ces méthodes de culture changent profondément la composition physico-chimique et biologique des terres. Par exemple, l’utilisation d’engrais enrichis en azote et en phosphore appauvrit la biodiversité tellurique des terres cultivées (Han et al., 2022; Wang et al., 2022).

Les méthodes de production déployées constituent également une source importante de pollution pour l’environnement. Par exemple, en Bretagne ou dans la baie du Mexique, les déchets issus de la production agricole sont rejetés dans l’océan. Cette pratique acidifie l’eau et favorise le développement d’algues qui asphyxient les écosystèmes (Bianchini et al., 2012; Landrigan et al., 2020).

Aujourd’hui, il n'y a plus de doute sur l'impact des méthodes de culture conventionnelle sur la santé humaine mais aussi sur la faune sauvage. Dès les années

80, l'analyse statistique de plusieurs cohortes d'hommes et de femmes confrontés à différents niveaux d'expositions aux substances chimiques utilisées en agriculture conventionnelle révèle, un risque accru de développement de cancers mais aussi de maladies neurologiques telles que la maladie de Parkinson (Fleming et al., 1999). Plusieurs analyses expérimentales ont mis en évidence les mécanismes moléculaires sous-jacents à ce type de maladies. Par exemple, Liu et collaborateurs ont montré que les produits chimiques de type carbamate, utilisés comme pesticides en agriculture conventionnelle, diminuent la viabilité cellulaire et induisent une déficience motrice chez la souris. L'analyse protéomique des tissus cérébraux de ces animaux indique une modification de voies métaboliques impliquées dans la maladie de Parkinson telles que les voies métaboliques de la phénylalanine et du tryptophane mais aussi du cycle des vésicules synaptiques (Liu et al., 2022). La prise de conscience de l'effet de ce mode de production sur la santé et les écosystèmes a remis en question ce mode de production et a favorisé la proposition de méthodes alternatives.

3. Développement des pratiques agricoles alternatives

Depuis les années 1980, des méthodes alternatives à l'agriculture conventionnelle basées sur les principes de l'agroécologie sont proposées. Cette notion regroupe un ensemble de théories et de pratiques agricoles inspirées par les connaissances de l'écologie, des sciences agronomiques et du monde agricole. Cette approche vise à concevoir des systèmes de production s'appuyant sur les fonctionnalités offertes par les écosystèmes (Altieri, 1999; Matson et al., 1997; Tscharntke et al., 2005).

Basée sur les principes de l'agroécologie, l'agriculture biologique représente, avec l'agriculture conventionnelle, l'un des deux modes de production présents en France et dans le monde. Selon l'article « Article L645-1 » du nouveau code rural, les méthodes de production ne doivent pas recourir à des produits chimiques de

synthèse. Cependant, l'éventail de méthodes proposées en agriculture biologique est encore faible et le rendement est moindre. De plus, des études doivent encore être menées pour évaluer l'efficacité de ces pratiques et leur impact environnemental (Rosenheim et al., 2022).

Suite à la révolution agricole, la population humaine n'a cessé d'augmenter et selon les modélisations démographiques, cette croissance devrait se poursuivre au moins jusqu'en 2050 et ainsi frôler les 10 milliards d'humains (Vollset et al., 2020). Selon certaines estimations, une augmentation de 25 à 70 % de la production agricole serait alors nécessaire pour subvenir aux demandes d'une telle population (Hunter et al., 2017). L'écart entre ces estimations repose sur l'impact de nombreux facteurs environnementaux et socio-économiques. Pour soutenir la demande alimentaire de manière équitable et durable et atteindre une sécurité alimentaire universelle, les schémas socio-économiques doivent profondément évoluer en termes de distribution et de consommation mais aussi de production. Pour minimiser les dommages collatéraux de la production agricole, il faut développer des pratiques efficaces et respectueuses de l'environnement mais aussi réduire les pertes liées aux attaques par des ravageurs de culture.

Pour répondre à ces enjeux, il est nécessaire d'avoir une compréhension globale des mécanismes biologiques intervenant dans les interactions des plantes avec leur environnement biotique et abiotique. Au niveau biotique, les plantes sont engagées dans des interactions qui peuvent être bénéfiques à leur santé et leur développement (symbiotes, commensaux) ou néfastes (parasites et pathogènes, compétiteurs). L'une des clés de l'agriculture de demain sera de favoriser les interactions bénéfiques en protégeant les cultures des interactions néfastes avec les organismes présents dans l'environnement de la plante. De plus, les études de microbiologie montrent que les communautés microbiennes jouent aussi un rôle très important sur la santé des plantes.

B. Évolution de la microbiologie

1. *Découverte des micro-organismes*

Bien que de nombreuses pratiques ancestrales, telles que la fermentation acétique utilisée depuis l'antiquité, sont basées sur les principes de la microbiologie, ce domaine de recherche a réellement vu le jour avec les progrès de la microscopie au XVIIème siècle (Finlay et Esteban, 2001). De nombreuses dates ont marqué l'histoire de la microbiologie, dont en particulier les travaux de recherche de Louis Pasteur portant sur la fermentation et la vaccination (Bernard et Pasteur, 1879; Pasteur, 1884).

Au XIXème siècle, Louis Pasteur démontre expérimentalement que les microbes sont partout, dans l'eau, dans l'air, sur les objets ou encore sur la peau. À l'époque, les micro-organismes étaient communément vus comme des « agents pathogènes » responsables de maladies conformément à la description initiale de Louis Pasteur. Cette vision « Pasteurienne » à forte connotation négative a persisté de nombreuses années, jusqu'à ce que les progrès technologiques révolutionnent la microbiologie. L'essor de la biologie moléculaire et des techniques de séquençage haut débit nous a permis d'avoir une meilleure compréhension du vivant, et plus particulièrement du monde microbien.

2. *Étendue et complexité du monde microbien*

Les communautés microbiennes sont aujourd'hui définies comme des assemblages multi-espèces, dans lesquels les micro-organismes interagissent les uns avec les autres dans un environnement contigu (Konopka, 2009). Le terme de micro-organismes regroupe, sur un critère de taille (visibilité), des espèces éloignées d'un point de vue phylogénétique. Ce terme désigne divers clades dont la plupart sont des Bactéries et des Archées mais aussi des virus et de nombreux organismes eucaryotes.

Les Bactéries et les Archées sont regroupées sous le terme de Procaryote, en raison de leur structure cellulaire dépourvue de noyau. Cependant, ce terme reposant sur une caractéristique morphologique (absence de noyau) et ne reflète pas un clade phylogénétique, comme nous le verrons par la suite. Il existe aussi des lignées microbiennes appartenant au clade des Eucaryotes, c'est-à-dire de micro-organismes constitués de cellules possédant un noyau. Parmi les Eucaryotes, les micro-organismes ont une distribution phylogénétique éparse qui constituent des groupes paraphylétiques.

3. *Distribution des micro-organismes dans l'arbre du vivant*

a) Connaissance partielle du monde microbien et classification à trois domaines

La classification du vivant était basée initialement sur des traits phénotypiques ou des caractères morphologiques. Les premières classifications du vivant étaient composées de deux domaines qui se rejoignaient à la racine de l'arbre de la vie (procaryotes/eucaryotes). Suite aux avancées en biologie moléculaire, les méthodes de classification ont été complétées par des caractères moléculaires insufflant un nouveau souffle à cette discipline. En 1990, ces progrès techniques et les avancées en microbiologie ont permis la découverte du clade des archées (Woese et al., 1990). Les auteurs de cette étude proposent alors un nouveau modèle de l'arbre du vivant, composé de trois domaines séparant les archées des bactéries au sein des procaryotes. Dans cette représentation, chacun de ces domaines correspond à un clade monophylétique distinct (Lake et al., 1984; Rivera et Lake, 1992). Lors de cette première étude, les archées apparaissaient déjà comme étant un groupe frère des eucaryotes et partagent un ancêtre commun unique distinct des bactéries (Figure 1 A).

Selon la théorie de l'endosymbiose propose d'expliquer l'évolution de la lignée eucaryote à partir d'un ancêtre procaryote, les cellules eucaryotes proviennent de l'incorporation d'une protéobactérie dans une cellule hôte dont l'identité est longtemps restée méconnue (Margulis, 1975; Martin et al., 2015). La protéobactérie qui aurait été incorporée représenterait aujourd'hui la mitochondrie des cellules eucaryotes actuelles. Selon la théorie de l'endosymbiose secondaire, les chloroplastes de cellules végétales proviendraient de l'incorporation successive de Cyanobactéries (Trench, 1975). Selon cette théorie, les organelles des cellules eucaryotes auraient été acquises par incorporation de cellules procaryotes.

b) Étude des micro-organismes 'non-cultivés' et redéfinition de l'arbre du vivant

Les avancées en microbiologie environnementale, permettant l'étude moléculaire de micro-organismes non cultivés, ont rapidement enrichi le catalogue de séquences biologiques disponibles. La découverte d'espèces de micro-organismes appartenant au domaine des archées (clades 'TACK' et 'Asgard') a redéfini notre compréhension des relations entre eucaryotes et archées (Brochier-Armanet et al., 2008). Les analyses cellulaires et la reconstruction des génomes de ces archées non-cultivées ont révélé la présence chez certaines espèces de caractéristiques spécifiques aux eucaryotes tel que des gènes codant pour l'actine. Les analyses phylogénétiques ont montré que ces nouvelles espèces d'archées sont étroitement apparentées aux Eucaryotes (Figure 1 B). L'intégration de ces nouvelles données et l'utilisation de modèles phylogénétiques plus fiables suggèrent que l'arbre de la vie est constitué de deux domaines principaux bactéries/archées, et les eucaryotes représentant une sous-branche des archées. Selon ces résultats (Figure 1 B), la lignée eucaryotes semble donc avoir émergé plus tardivement à partir du clade des archées (Eme et al., 2017; Spang et al., 2015; Williams et al., 2020; Zaremba-Niedzwiedzka et al., 2017).

Cela montre que notre compréhension du vivant est conditionnée par les progrès technologiques. Récemment, l'essor de la biologie moléculaire et des techniques de séquençage haut débit ont permis de révolutionner le domaine de la microbiologie grâce au développement d'approches métagénomiques qui permettent d'étudier le contenu moléculaire des communautés microbiennes.

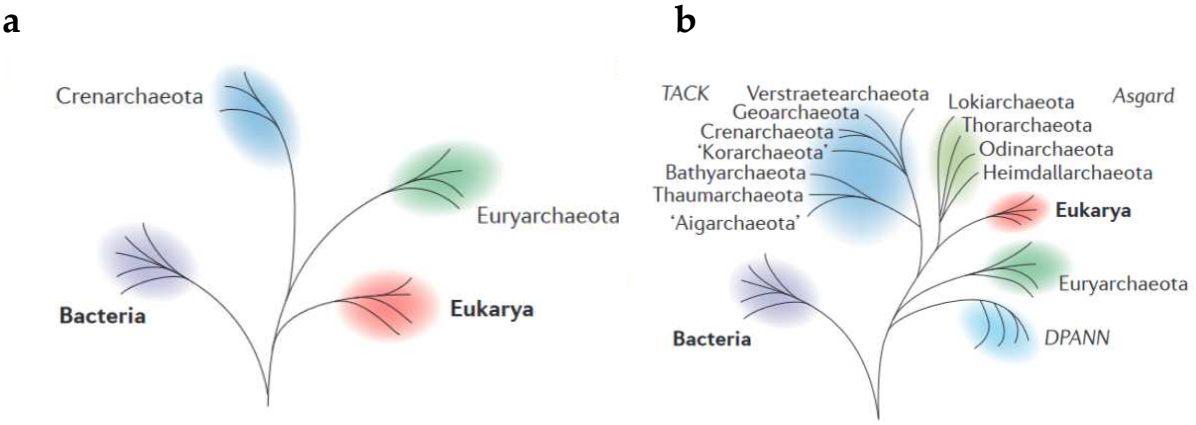


Figure 1. - Représentations schématiques de l'arbre phylogénétique du vivant (A) à trois domaines comprenant les bactéries, les eucaryotes (Eukarya) et les archées (Crenarchaeota et Euryarchaeota) représentant chacun un groupe monophylétique partageant un ancêtre commun unique distinct des bactéries. Suite à la découverte de membres du superphylum Asgard appartenant aux archées, ce modèle a ensuite été remplacé par un arbre à deux domaines ou les eucaryotes ont évolué à partir du clade des Asgard. Dans ce schéma, la branche correspondant aux bactéries est représentée en violet, les Eukarya sont de couleur rouge et les lignées d'archées sont représentées en vert et bleu. Figures issues de (Eme et al., 2017).

c) La métagénomique pour étudier les communautés microbiennes 'non-cultivées'

Introduit pour la première fois en 1998 dans un travail portant sur les micro-organismes telluriques, le terme « métagénome » désigne l'ensemble des génomes des micro-organismes contenus dans un milieu (Handelsman et al., 1998).

(i) Développement des approches métagénomiques

Pendant longtemps, la majorité des données génomiques microbiennes disponibles provenaient de souches cultivées en laboratoire. Comme mentionné par Handelsman et collaborateurs, la majorité des micro-organismes présents dans les sols ne peuvent pas être cultivés (Handelsman et al., 1998). Pour explorer la diversité génétique et le potentiel fonctionnel des communautés microbiennes, il était nécessaire de développer une méthode proposant de travailler directement sur l'ADN extrait d'un milieu naturel. Dans cet article, Handelsman et collaborateurs proposaient le clonage du métagénome afin d'isoler de nouvelles voies de synthèse de molécules bioactives à partir de micro-organismes du sol. Les procédés reposent sur l'extraction de l'ADN à partir d'échantillons de sol. Cette extraction est suivie d'un découpage à l'aide d'enzymes de restriction puis clonage des fragments dans un chromosome artificiel bactérien (BAC). Dans cette méthode, les clones BAC étaient ensuite testés pour leur activité biologique et pour la production de nouveaux produits naturels. Ces méthodes fastidieuses sont maintenant appuyées par les technologies de séquençage apparues dans les années 2000.

Les approches métagénomiques sont basées sur le séquençage de la globalité de l'ADN extrait d'un échantillon de microbiome. Contrairement à l'approche génomique qui vise à étudier l'ensemble des gènes d'un organisme, la métagénomique étudie l'ensemble des gènes des micro-organismes d'un milieu appelé microbiome.

(ii) Distinction Microbiome/Microbiote

En 1988, Whipps et ses collaborateurs qui travaillaient sur l'écologie des micro-organismes de la rhizosphère (la partie du sol entourant les racines) ont fourni la première définition du terme microbiome. Ils ont décrit le « microbiome » comme une « communauté microbienne caractéristique » d'un « habitat raisonnablement bien défini qui présente des propriétés physico-chimiques distinctes », et son «

théâtre d'activité » qui réfère aux conditions physico-chimiques mais aussi à toutes les molécules présentes dans ce milieu. Depuis la proposition de cette définition par Whipps, de nombreuses autres définitions du microbiome ont été formulées mais celle-ci reste la plus pertinente car elle relie les notions de communauté microbienne et de spécificité de niche écologique. La majorité des autres définitions ne rendent pas compte de la complexité de ce terme. En effet, il est souvent réduit à la même définition que le microbiote qui fait uniquement référence aux « acteurs » présents sur cette « scène ». La définition de Whipps & collaborateurs a d'ailleurs été reprise dans le cadre de l'initiative *MicrobiomeSupport*, financée par l'Europe, dont l'un des objectifs était de définir un lexique consensus pour unifier les concepts de ces domaines (Berg, 2020).

Par conséquent, la métagénomique permet d'étudier l'ensemble des génomes et fragments d'ADN contenus dans un microbiome. Cette approche permet d'envisager l'étude de tous microbiomes accessibles, à partir desquels de l'ADN peut-être extrait et purifié, tout en conservant l'intégrité des molécules. On distingue classiquement les microbiomes environnementaux (eaux, sols, air) des microbiomes associés à un hôte (animaux, plantes).

(iii) Grandes étapes de l'analyse métagénomique

Parmi les approches métagénomiques, on distingue le séquençage *shotgun* qui consiste à séquencer l'ensemble de l'ADN extrait d'un microbiome, du séquençage ciblé de courts fragments d'ADN marqueur appelé *metabarcoding*.

Après échantillonage, la première étape d'une analyse métagénomique « shotgun » est l'extraction de l'ADN présent dans le milieu microbien prélevé selon un protocole spécifique à la technologie de séquençage choisie. Puis, l'échantillon est chargé sur un séquenceur qui va « lire » les séquences nucléotidiques (composées d'Adénine, Thymine, Guanine et Cytosine), et fournir un ensemble de lectures de ces séquences sous forme de lettres correspondant aux nucléotides (respectivement A, T,

G et C). Selon les technologies de séquençage utilisées, la tailles des lectures peut varier de 150 pb (Illumina) à 800 kb (Oxford Nanopore). A partir de ces lectures, il est possible d'envisager de reconstruire des « Metagenomes Assembled Genomes » (i.e les génomes ou fragments de génomes). Cette étape de reconstruction constitue un véritable défi méthodologique, compte tenu de la difficulté d'assembler indépendamment les génomes à partir du mélange de lectures obtenu par la majorité des méthodes de séquençage. Ce problème est exacerbé si les lectures disponibles sont courtes ou très bruitées en raison de la technique de séquençage choisie ou de la qualité des échantillons. Des méthodes de binning sont souvent nécessaires pour regrouper les fragments selon leur origine, et l'assemblage reste souvent partiel.

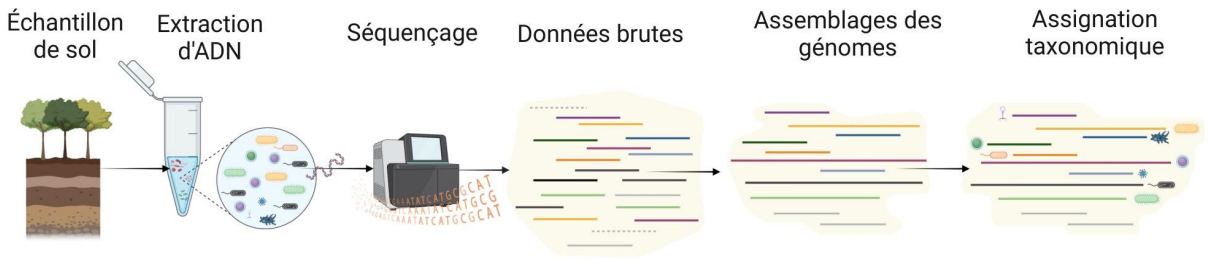


Figure 2. Les principales étapes d'une analyse métagénomique shotgun.

L'assemblage des métagénomes permet d'étudier la composition moléculaire et structurale des génomes ou encore d'effectuer des prédictions des gènes pour découvrir le potentiel fonctionnel des individus présents dans l'échantillon (Figure 2). La reconstitution de génomes de bonne qualité permet d'envisager une assignation taxonomique par placement phylogénétique (Chaumeil et al., 2020) .

(iv) Redéfinition du monde microbien par la métagénomique

Bien que certains verrous méthodologiques subsistent, les approches métagénomiques ont déjà permis de franchir un énorme pas dans la compréhension

du vivant et des relations entre règnes. Cela a provoqué un profond changement de notre vision du monde microbien. Le terme « microbe » a été progressivement remplacé par celui de micro-organisme faisant nettement évoluer notre vision autant en santé humaine qu'en agronomie.

On sait depuis plusieurs décennies que certains micro-organismes ont une relation directe avec l'être humain et qu'on peut les retrouver dans divers organes, ou encore dans certains procédés alimentaires. Cependant, en raison des limitations techniques, les relations étudiées étaient principalement néfastes, conduisant à une vision manichéenne des organismes microbiens. La métagénomique a permis d'étudier plus largement et sans *a priori* la diversité des micro-organismes et des rôles qu'ils jouent dans les écosystèmes. Nous avons maintenant une représentation plus complète du vivant, provoquant un changement de paradigme vers une vision intégrative où les micro-organismes contribuent à l'équilibre des écosystèmes.

Ainsi, nous sommes passés du concept de micro-organisme comme cellule unique et indépendante au concept d'entité participant à un assemblage complexe où les interactions et communications sont essentielles. Sur la base de ces observations, la théorie de l'holobionte a émergée. Le terme holobionte, proposé par Lynn Margulis, fait référence à un « superorganisme » ou « méta organisme » théorique qui comprend un organisme animal/végétal et les micro-organismes qu'il héberge (Margulis, 1990). Désormais, il est acquis que tout organisme eucaryote pluricellulaire doit être considéré avec son microbiote comme une unité fonctionnelle indissociable.

Un déséquilibre de la composition (i.e. diversité d'espèces) d'un microbiote peut entraîner un phénomène en cascade aboutissant à l'apparition de pathogènes. Ce phénomène appelé dysbiose, découvert en santé humaine dans le cadre des nombreux travaux sur le microbiome intestinal (Stecher et al., 2013), est aussi observé en santé des plantes au niveau de la rhizosphère et du microbiome terrestre

(Dastogeer et al., 2022). Ces connaissances ont fourni une nouvelle vision qui a influencé l'adoption de comportements plus « conservateurs » envers la diversité des micro-organismes, autant en santé humaine qu'en agronomie. Aujourd'hui, les produits en relation avec le microbiote du sol représentent le secteur le plus dynamique de l'agronomie.

C. Rôle des communautés telluriques dans la santé des plantes

Le sol représente une importante surface d'échange et d'interaction de la plante avec son environnement. Pour les pédologues, le sol est une surface meuble d'épaisseur variable résultant de la transformation de la roche mère sous jacente sous l'influence de divers processus physiques, chimiques et biologiques et sur laquelle peut se développer un végétal (Demelon, 1960 & Lassinier, 1973). Lieu de vie de nombreux organismes, cette matrice présente une riche diversité animale, végétale et microbienne. On considère comme vivant dans les sols, les organismes dont au moins une partie du cycle de vie se déroule dans cette matrice.

Les micro-organismes des sols, ou telluriques, ont un rôle essentiel dans la biosphère terrestre, la santé humaine, la biogéochimie mondiale, l'influence de la fertilité des sols et les échanges de CO_2 ou autres gaz à effet de serre (Jansson et Hofmockel, 2020). Cette biodiversité terrestre est composée de communautés microbiennes présentant des structures flexibles et des relations complexes modulables qui peuvent atténuer les effets des modifications biotiques et abiotiques sur la plante (Bonkowski, 2004; Crossay et al., 2019; Schouteden et al., 2015; Topalović et al., 2022). Cependant, la biodiversité du sol reste globalement méconnue, autant sur le plan écologique que taxonomique car la plupart des organismes constituant le microbiote du sol ne sont pas cultivables en laboratoire et donc difficilement étudiables (Handelsman et al., 1998). Le sol est considéré comme

la « troisième frontière biotique » après les grands fonds océaniques et les canopées, c'est-à-dire l'un des milieux dont l'homme est encore loin d'avoir exploré toute la richesse du fait de sa complexité et de son impressionnante diversité (André et al., 1994).

On estime que le microbiote du sol présente une très forte diversité. Il existerait ainsi plus de deux millions d'espèces de bactéries et de champignons dont moins de 1 % sont cultivés en laboratoire (Torsvik et Øvreås, 2002). Bien que les approches métagénomiques développées ces dernières décennies permettent d'explorer cette diversité, notre connaissance de ces communautés microbiennes telluriques reste encore très partielle (Fierer, 2017). L'étude des micro-organismes telluriques est particulièrement limitée par la difficulté à conserver l'intégrité des molécules d'ADN durant les phases d'extraction et de purification. Malgré l'optimisation des protocoles, les fragments d'ADN obtenus restent souvent fragmentés, de nombreuses molécules persistantes dans les échantillons peuvent limiter l'étape de séquençage et les méthodes bio-informatiques ne permettent qu'une reconstruction partielle des génomes.

Outre l'ensemble de micro-organismes, la rhizosphère comprend aussi les organes souterrains des végétaux incluant les racines des plantes qui représentent, en termes de surface, la moitié de l'interface de la plante avec l'environnement. La rhizosphère est le lieu de vie de nombreux micro-organismes qui vont pouvoir entrer en interaction avec la plante. Il s'agit donc d'une zone très importante pour le développement et pour la santé de la plante. Parmi les relations qui vont contribuer à la santé de la plante, on distingue les relations mutualistes (relation non obligatoire) des relations symbiotiques (obligatoire à la survie de l'un des partenaires). D'autre part, certaines interactions biotiques ont un effet néfaste qui altère la santé de la plante. Certaines théories, telle que celle de l'effet Janzen-Connell, postulent que les interactions parasitaires sont, d'un point de vue global, bénéfiques pour la biodiversité d'un écosystème. Selon cette théorie, quand

des micro-organismes pathogènes sont présents, la pousse de plantules de l'espèce sensible est inhibée, laissant la place à d'autres espèces de plantes à proximité (Clark et Clark, 1984). En modulant les densités de population, ce phénomène favorise donc la diversité végétale dans les espaces naturels (sous réserve d'une virulence modérée). Les pathogènes, en régulant les populations, jouent ainsi un rôle essentiel dans l'équilibre des écosystèmes. Cependant, dans un espace agricole, les phytopathogènes diminuent les rendements, ou pire, anéantissent les récoltes. En effet, les pertes agricoles causées par les parasites et pathogènes sur les cinq grandes cultures au niveau mondial, que représentent le blé, le riz, le maïs, la pomme de terre et le soja, s'élèvent en moyenne entre 17 et 30% de la production (Savary et al., 2019). Une grande part des dommages agricoles sont causés par des nématodes parasites de plantes qui englobent à eux seuls environ 12% des pertes mondiales (Singh et al., 2015).

2. *Les nématodes phytoparasites*

A. Les nématodes

1. *Description générale*

Les nématodes, ou vers ronds, représentent un embranchement monophylétique englobant l'ensemble des espèces de vers non segmentés. Ils appartiennent au taxon des Ecdysozoaires qui inclut aussi les arthropodes (i.e. insectes, crustacés, arachnides) ou les tardigrades (Figure 3). Ce taxon regroupe les animaux dont le développement implique des mues successives (Borner et al., 2014).

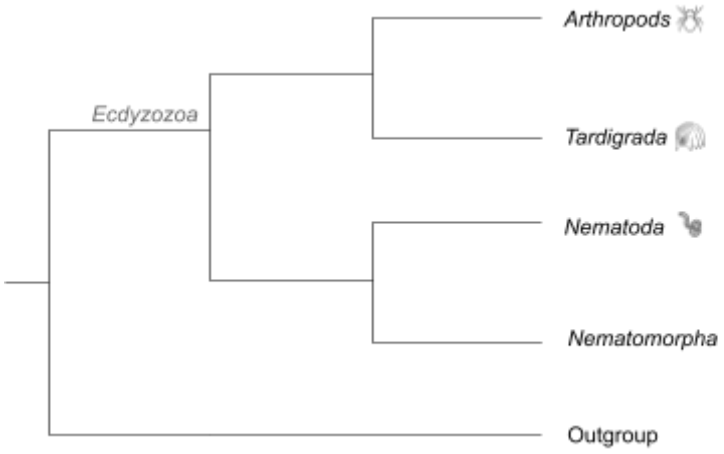


Figure 3. Représentation schématique de l'arbre phylogénétique du clade « Ecdysozoa » réalisé à partir du travail de (Yamasaki et al., 2015). Arbre obtenu par maximum de vraisemblance à partir de données 18S et 28S.

Les nématodes sont des organismes vermiformes avec une longueur généralement comprise entre 5 et 100 µm pour un diamètre de 20 µm. La plupart sont translucides et invisibles à l'œil nu. Certaines espèces qui ont un mode de vie libre peuvent atteindre une taille de 5 cm et d'autres espèces parasites d'animaux peuvent atteindre une longueur de plusieurs mètres. Le record en la matière étant *Placentonema gigantissima*, parasite du placenta des cachalots pouvant dépasser les 8 mètres de longueur.

2. *Distribution géographique*

Les analyses microscopiques ont permis de mettre en évidence la présence de nématodes dans tous les milieux (i.e. terrestres, marins, d'eau douce) en association avec des animaux ou bien des plantes (Bardgett et van der Putten, 2014; Ferraz et Brown, 2002; van den Hoogen et al., 2019). On les retrouve sur tous les continents avec des densités de population variant selon la latitude. Étonnamment, les populations de nématodes sont plus denses dans les régions froides comme la Russie ou le Canada que dans les régions chaudes telles qu'en Afrique du nord (Figure 4 A). Cependant, l'étude étant limitée au sol proprement dit, elle pourrait sous-estimer l'abondante diversité de nématodes présents au niveau de la litière dans les forêts tropicales selon Hoogen et collaborateurs.

3. *Modes de vie, types trophiques et rôle écologique*

La plupart des espèces sont bénéfiques pour l'agriculture. Elle participent à la décomposition de la matière organique et au cycle trophique et elles jouent ainsi un rôle essentiel dans la chaîne alimentaire du sol. Tous les types trophiques retrouvés chez les animaux sont représentés chez les nématodes. Il existe des espèces bactérivores, fongivores, omnivores et même prédatrices d'autres nématodes qui ont un mode de vie libre. Avec des espèces très diversifiées, ils occupent des fonctions essentielles dans les réseaux trophiques terrestres. Le type trophique le plus abondant est celui des bactériophages qui comprend plusieurs espèces vivant librement et représentant la majorité de la biomasse totale de nématodes sur terre (Figure 4 B). Ces nématodes peuvent jouer des rôles importants dans les écosystèmes mais n'interagissent pas de manière directe avec les autres organismes contrairement à d'autres espèces dites « parasites ». Il existe de nombreuses espèces de nématodes parasites d'animaux (principalement d'insectes et de vertébrés) ou de plantes. L'étude réalisée par Hoogen et collaborateurs montre que les espèces parasites de plantes constituent le second type trophique le plus dense parmi les nématodes,

comme indiqué dans la figure 4 où les NPP apparaissent sous le terme « herbivore » (Figure 4 B).

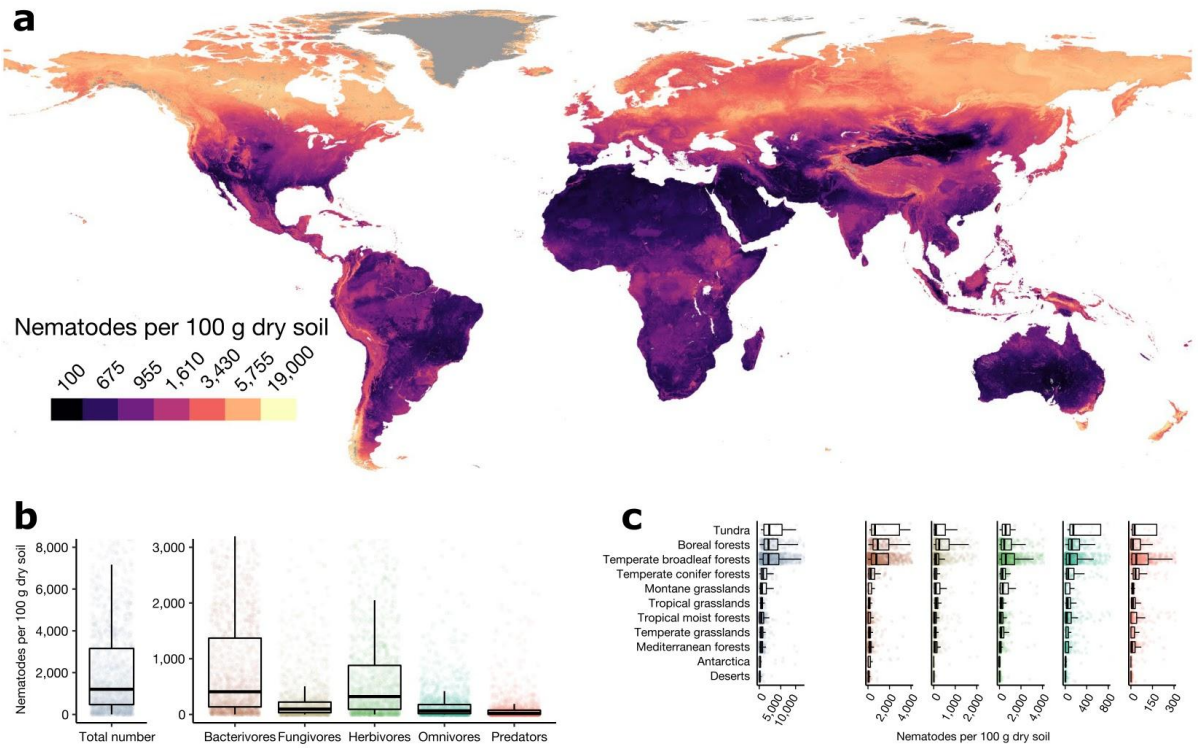


Figure 4. Informations sur la distribution mondiale des nématodes issus de l'analyse de 6 759 échantillons géo-référencés. (A) Carte mondiale de la densité des nématodes dans le sol. (Nombre de nématodes par 100 g de sol sec). (B) Abondances de nématodes (n = 1 876) par groupe biotope de tous les continents. (C) La médiane et l'écart interquartile des abondances de nématodes (n = 1 876) par groupe trophique et par biotope. Figure modifiée issue de (Van den Hoogen et al., 2019).

B. Les nématodes parasites de plantes

1. *Description générale*

En raison de leur taille microscopique, la première description d'un nématode parasite de plantes a seulement été rapportée en 1879 par Cornu et concernait des nématodes trouvés dans des galles de racines de sainfoin (*Onobrychis sativus Lam.*) dans la vallée de la Loire, en France. Quelques années plus tard, un nématode à galles provenant de caféiers au Brésil a été brièvement décrit, illustré et nommé *Meloidogyne exigua* par Göldi (Perry et al., 2009).

Aujourd'hui, plus de 4 000 espèces de nématodes ont été identifiées comme parasites de plantes, ce qui correspond à 7% du phylum Nematoda (Decraemer et Hunt, 2006). On les retrouve principalement dans les zones tempérées (Figure 4 C) telles que les zones de forêt tempérée de feuillus. Leur distribution géographique est corrélée à celle des espèces végétales en raison du statut de parasite obligatoire de nombreuses espèces dont la prolifération dépend de la présence de la plante hôte (Hoogen, 2019).

Il existe des espèces de nématodes capables d'infecter la majorité des végétaux dont de nombreuses espèces d'intérêt agronomique telles que la tomate, le riz, la pomme de terre ou le soja. L'ensemble des espèces de NPP représente une contrainte importante pour la sécurité alimentaire à l'échelle mondiale et le développement de méthodes de lutte durables est une priorité.

2. *Impact agricole des nématodes parasites de plantes*

Les NPP représentent un stress biotique important pour les végétaux. Ils impactent le fonctionnement normal de la plante en détournant leurs ressources, ce qui provoque une diminution du rendement. Ainsi, les pertes agricoles causées par les nématodes phytoparasites s'élèvent en moyenne à 12,3% de la production mondiale, ce qui correspond à plus de 157 milliards de dollars de pertes par an (Abad et al., 2008; Singh et al., 2015). Il est fort possible que ces valeurs soient sous-estimées car ces nématodes sont généralement de petits agents pathogènes transmis par le sol et les symptômes qu'ils provoquent sont souvent non spécifiques et difficiles à diagnostiquer.

La majorité des cultures sont susceptibles d'être infectées par les NPP et ils sont retrouvés dans toutes les régions agricoles du monde. Certaines régions sont plus ou moins touchées et, à l'échelle mondiale, la distribution des espèces de NPP varie considérablement. Par exemple, en Inde, les pertes relatives aux NPP s'étendaient à 21,3% en 2020, soit 1,58 milliard de dollars uniquement pour ce pays

(Kumar et al., 2020). Les espèces les plus critiques peuvent également varier selon les régions. Les connaissances des facteurs environnementaux et agronomiques ainsi que de la diversité des espèces permettent de comprendre la distribution de ces organismes. Ces éléments présentent une importance capitale dans le développement de méthodes de luttes spécifiques. Les cartes de distribution et les données sur la gamme d'hôtes de certaines espèces sont disponibles et régulièrement mises à jour. Elles constituent une source utile pour déterminer le potentiel de dommages causés par les nématodes (http://www.cabi.org/dmpd).

Certaines espèces sont cosmopolites. C'est notamment le cas de quatre espèces de nématodes à galles du genre *Meloidogyne* : *Meloidogyne incognita, Meloidogyne arenaria, Meloidogyne javanica et Meloidogyne hapla* (Subbotin et al., 2021). D'autres espèces sont particulièrement limitées géographiquement comme c'est le cas pour 49 espèces décrites de *Meloidogyne* qui n'ont été enregistrées jusqu'à présent que dans leur localité typique. Par exemple, *Meloidogyne dunensis* a été identifié uniquement en Europe ou encore *Meloidogyne aquatilis* en Amérique du Nord (Subbotin et al., 2021).

En 2013, John T. Jones et collaborateurs ont établi, à l'occasion d'une revue dans le journal *Molecular Plant Pathology,* la liste du 'top 10' de ces pathogènes en fonction de leur importance scientifique et économique :

(1) les nématodes à galles (*Meloidogyne spp.*); (2) les nématodes à kyste (*Heterodera et Globodera spp.*); (3) les nématodes des lésions racinaires (*Pratylenchus spp.*) ; (4) le nématode fouisseur *Radopholus similis*; (5) *Ditylenchus dipsaci*; (6) le nématode du flétrissement du pin *Bursaphelenchus xylophilus*; (7) le nématode réniforme *Rotylenchulus reniformis*; (8) le nématode vecteur de virus *Xiphinema index* ; (9) *Nacobbus aberrans*; et (10) *Aphelenchoides besseyi* (Jones et al., 2013). Les nématodes à galles (*Meloidogyne)*, et plus spécifiquement *M. incognita,* sont responsables de la majorité des pertes en tabac et tomate, comme *M. javanica* pour le tournesol et le poivron (Wesemael et al, 2011).

Ainsi, les espèces du genre *Meloidogyne* sont parmi les parasites de plantes décrits comme les plus grands ravageurs de culture. Leur impact économique est ainsi très important en raison de leur large gamme d'hôtes et de leur distribution dans les environnements tempérés et tropicaux (Perry et al., 2009). Les nématodes phytoparasites sédentaires comme les nématodes à galles et à kystes développent des interactions compatibles avec un large éventail de plantes cultivées, notamment le blé (*Triticum aestivum L.*), la pomme de terre (*Solanum tuberosum L.*), la tomate (*S. lycopersicum L.*), le soja (*Glycine max (L.) Merr.*) et la betterave sucrière (*Beta vulgaris L.*). Le nématode à galles, *Meloidogyne incognita* est capable d'infecter à lui seul plus de 3 000 espèces végétales dont de nombreuses plantes cultivées (Kofoid et White, Abad et al., 2008).

De nombreux NPP sont des parasites polyphages qui peuvent infecter toutes sortes de cultures. Les dégâts qu'ils engendrent peuvent provenir directement de leur pénétration et migration dans les tissus de la plante, du détournement du métabolisme de la plante mais aussi de l'infection secondaire par d'autres pathogènes comme des champignons ou des virus (Garcia et al., 2022; Taylor et Robertson, 1970).

3. *Méthodes de luttes contre les nématodes parasites de plantes*

a) Les méthodes conventionnelles

Actuellement, il n'existe pas de méthode efficace et durable pour faire face à ces ravageurs de cultures. Les nématicides utilisés en agriculture conventionnelle sont principalement composés de molécules chimiques appartenant aux familles des carbamates, des hydrocarbures halogénés ou des organophosphorés. Dans la majorité des pays, ces substances sont interdites car elles ont des effets néfastes sur la santé humaine et l'environnement (King et Aaron, 2015; Nicol et al., 2011). L'utilisation de produits chimiques non spécifiques est aussi néfaste pour la biodiversité des sols (Rahman et Zhang, 2018), et la productivité à long terme (Seenivasagan et Babalola, 2021). Depuis une dizaine d'années en Europe, les

stratégies de gestion des infections par les NPP doivent respecter la directive sur la lutte intégrée contre les parasites de l'Union européenne, qui vise à réduire l'usage des pesticides et à promouvoir autant que possible les pratiques de gestion non chimiques (*directive 200//128/CE*).

La réduction de l'utilisation de produits chimiques laisse place au développement d'infections devant être contenues par des méthodes alternatives. De plus, les changements climatiques pourraient influencer la fréquence, l'intensité et la distribution des infections par les NPP. Ces changements vont modifier le cycle de développement du parasite (Okulewicz, 2017) mais aussi favoriser la colonisation de nouveaux espaces (Bebber et al., 2013). Ainsi, une augmentation des dommages causés par les NPP est à craindre dans les prochaines années. Dans ce contexte, des méthodes alternatives aux nématicides ont été proposées durant les dernières décennies.

b) Les méthodes alternatives

(i) Optimisation de la composition des communautés microbiennes

Comme mentionné précédemment, une partie substantielle des organismes présents naturellement dans les sols sont susceptibles d'avoir un effet bénéfique direct ou indirect sur la santé des plantes tel que certaines espèces de bactéries ou champignons. Ainsi, des approches alternatives reposent sur l'optimisation de la composition des communautés microbiennes telluriques.

(ii) Les micro-organismes bénéfiques à la santé des plantes

Les micro-organismes qui entretiennent une relation symbiotique avec les plantes ont un effet bénéfique sur la santé et la croissance de la plante, mais aussi sur la réduction des attaques des pathogènes et des parasites. C'est le cas, par exemple, des champignons mycorhiziens à arbuscules (CMA) appartenant à la classe des *Gloméromycètes*, qui favorisent la croissance via l'apport en nutriments tel que l'azote

et le phosphore (Estrada et al., 2013; Hestrin et al., 2022; Liu et al., 1997; Sellitto et al., 2019; Taoheed et al., 2018), et la tolérance de la plante aux stress biotiques tels que les infections par les NPP (Detrey et al., 2022; Schouteden et al., 2015). Les auteurs suggèrent qu'en présence de CMA, il y a une compétition directe entre les NPP et les CMA pour l'accès aux sites de nutrition, une tolérance accrue des plantes aux différentes sources de stress et une réduction des exsudats racinaires, ce qui modifie les interactions de la plante avec les autres membres de la rhizosphère (Bell et al., 2022).

Sur la base de ces observations, des méthodes reposant sur la valorisation de la synergie microbiome - plante sont proposées dans la littérature scientifique (Chen et al., 2018) et se développent sur le terrain via la commercialisation de solutions mycorhiziennes telles que proposées par la société Mycophyto. Cependant, les effets de l'interaction plante-CMA sont hautement dépendants des conditions environnementales (Balzergue, 2012; Candido et al., 2013; Nedorost et Pokluda, 2012) mais aussi de l'identité de la plante d'intérêt (Alguacil et al., 2018; L. Chen et al., 2017; Eom et al., 2000; Garzo et al., 2020).

(iii) Les micro-organismes antagonistes des NPP

Il existe aussi des approches alternatives utilisant des micro-organismes ciblant spécifiquement les NPP. Par exemple, certaines bactéries et des champignons ont été identifiés comme antagonistes de ces nématodes, c'est le cas des espèces fongiques *Arthrobotrys irregularis* par exemple. Ces champignons sont des prédateurs des espèces du genre *Meloidogyne* (Cayrol, J.C., 2013), ou *Paecilomyces lilacinus*, ovicide qui tue les œufs des espèces de nématodes à galles mais aussi d'espèces de nématodes à kystes (Gomes Carneiro et Cayrol, 1991). Cependant, les méthodes proposées jusqu'ici nécessitent une mise en œuvre assez lourde pour une efficacité qui reste partielle. Pour optimiser le développement et la mise en place de ce type de méthodes, il est essentiel de mieux comprendre les relations intervenant entre les différents membres de ces communautés.

(iv) Utilisation de plantes résistantes et rotation des cultures

Une autre méthode de lutte repose sur l'utilisation de plantes résistantes aux NPP. Depuis de nombreuses années, on sait que certaines variétés végétales peuvent être naturellement résistantes aux pathogènes ou parasites comme les nématodes (Hare, 1956; Hendy et al., 1985). Chez de nombreuses espèces végétales, cette résistance est liée à la présence de gènes appelés « gènes de résistance ». Par exemple, chez le piment (*Capsicum annuum*), la résistance aux NPP repose sur six gènes indépendants qui confèrent une résistance à certaines espèces de nématodes de manière plus ou moins spécifique (Djian-Caporalino et al., 2019).

Ces gènes de résistance engendrent différents schémas de réponse dans les cellules racinaires en fonction de la lignée végétale et de l'espèce de nématode. L'un des mécanismes couramment observé est l'induction d'une nécrose des tissus environnant le site d'infection, qui permet de contenir la propagation des nématodes en limitant leur dispersion et leur reproduction.

On retrouve généralement ces phénotypes résistants chez des espèces sauvages qui ne présentent pas d'intérêt agronomique. Cependant des processus de croisement et de sélection permettent de générer des variétés résistantes d'intérêt agronomique. Une autre technique consiste à générer des plantes résistantes par clonage ou semis puis à les décapiter pour greffer la partie aérienne de la plante d'intérêt. Sous réserve de trouver un couple compatible, le développement de ce type de système est plus rapide que la procédure de sélection qui peut prendre une dizaine d'années. L'utilisation de pieds porte-greffe est très répandue comme par exemple pour la culture de la vigne.

L'utilisation de gènes de résistance permet de réduire le pouvoir infectieux d'un sol mais elle limite le choix de variétés cultivables. De plus, l'efficacité est souvent partielle et le mécanisme peut-être contourné au cours du temps par les

nématodes. Ces approches doivent être combinées avec d'autres stratégies comme la rotation des cultures. La rotation des cultures, ou assolement, est souvent difficile à gérer et contraignante pour l'agriculteur. Aussi, l'efficacité est limitée car certaines espèces, telles que *Heterodera avenae* ou *Heterodera filipjevi* (espèces de nématodes à kyste parasitaires de la pomme de terre), sont résistantes et peuvent survivre sous une forme dormante plusieurs mois voire plusieurs années.

Une autre approche est la lutte intégrée contre les ravageurs. Il s'agit d'une approche à grande échelle qui vise à intégrer à la fois des pratiques chimiques et non chimiques pour un contrôle optimal des ravageurs. L'objectif de cette approche est de diminuer les populations de ravageurs en dessous du niveau préjudiciable. Ce n'est pas une simple juxtaposition ou superposition de deux techniques de lutte (telles que la lutte chimique et la lutte biologique) mais l'intégration de toutes les techniques de lutte adaptées. Cette approche offre un bon compromis entre efficacité et impact mais elle reste très difficile à mettre en œuvre et demande une bonne compréhension des systèmes agricoles et naturels.

Malgré les dommages causés par les nématodes, les mécanismes moléculaires utilisés par ces organismes pour acquérir leur virulence restent largement inconnus. Pour développer de nouvelles méthodes de lutte intégrées contre ces parasites et être en mesure de déployer ce type d'approches, il est essentiel de mieux comprendre la biologie de ces organismes et ces mécanismes moléculaires.

4. *Classification phylogénétique des nématodes parasites de plante*

On observe chez les nématodes parasites de plantes certaines similarités morphologiques (par exemple la présence d'un stylet) et biologiques (par exemple la sécrétion d'effecteurs du parasitisme). Cependant, les stratégies développées et les processus moléculaires impliqués sont très diversifiés et résultent d'événements d'adaptation évolutive indépendants et convergents (Figure 5) (Quist et al., 2015).

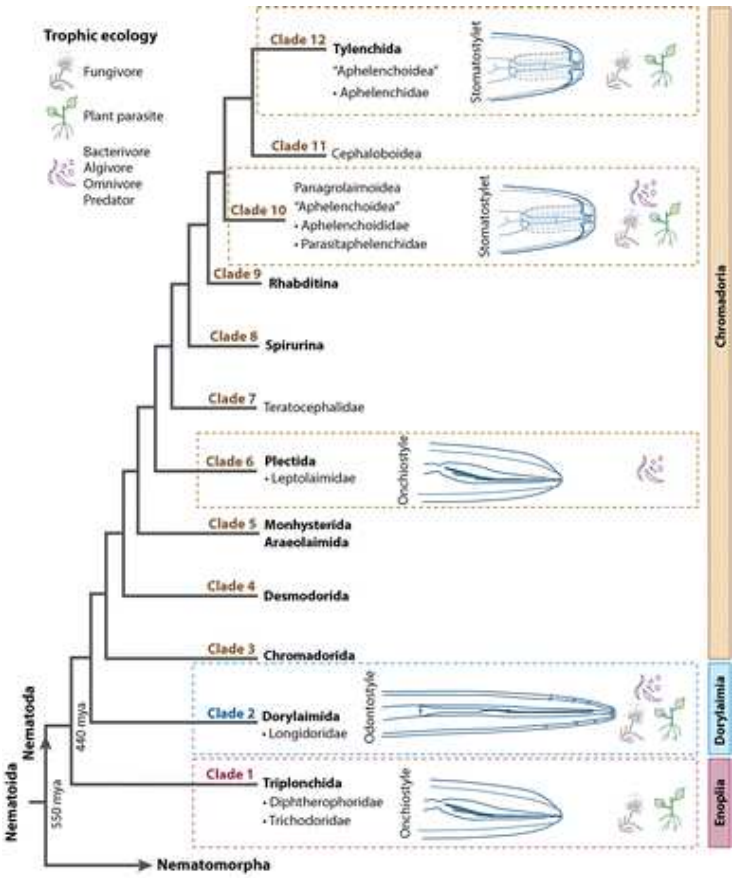


Figure 5. Une vue d'ensemble schématique de la division du phylum nematoda en douze clades majeurs est basée sur les séquences SSU rDNA selon la phylogénie de Van Megen. Présentation de différents types de stylets de nématodes, dont le stomatostylet, l'odontostylet et l'onchiostylet, respectivement dans les clades 12, 2 et 1. Figure issue de (Quist et al., 2015)

La première classification phylogénétique des nématodes, proposée par Chitwood, était basée sur des caractères morphologiques tels que la présence de glandes et/ou d'organes sensoriels (Chitwood, 1958). Les nématodes arborent un nombre limité de caractères morphologiques distinctifs difficiles à distinguer (Holterman et al., 2017) et l'identification d'espèces sur la base de ces caractères requiert une grande expertise.

Le développement des techniques de séquençage haut débit a permis d'intégrer dans les classifications phylogénétiques des caractères moléculaires et d'étendre l'analyse à un plus grand nombre d'espèces. Les nématodes étant probablement apparus au début du Cambrien (550 millions d'années), seuls les

gènes hautement conservés peuvent être utilisés pour reconstruire l'histoire phylogénétique globale de ce phylum. Ainsi, les premières phylogénies moléculaires étaient basées sur les variations de séquences codant pour les petites sous-unités ribosomales, un gène non impliqué dans la pathogénicité. Dès ces premières classifications, les NPP sont apparus comme polyphylétiques, c'est-à-dire possédant un caractère similaire qui n'a pas été hérité d'un unique ancêtre commun. Ces résultats suggèrent de multiples émergences indépendantes du phyto-parasitisme dans le phylum des nématodes (Blaxter et al., 1998; Liu et al., 1997). Dans la phylogénie de Blaxter et al., publiée en 1998, les NPP sont distribués dans trois des cinq taxons « majeurs » du phylum Nematoda ressortant de cette étude (Figure 6).

Depuis ces travaux, de nombreuses autres reconstructions phylogénétiques ont été réalisées pour ce phylum. Les analyses englobent de plus en plus d'espèces et de caractères afin d'obtenir une résolution plus fine (Blaxter et Koutsovoulos, 2015; De Ley, 2006; Holterman et al., 2017, 2009, 2006; van Megen et al., 2009). Parfois, les résultats obtenus varient sensiblement selon les caractères, les espèces et les protocoles utilisés. Par exemple, la phylogénie de Holterman et van Megen, basée sur les séquences de la petite sous-unité de l'ADN ribosomal identifie une douzaine de clades dans le phylum Nematoda. Au moins quatre de ces clades (clade 1,2, 10 et 12) comprennent des nématodes parasites de plantes (Figure 6) :

- Les clades 1 et 2 sont des groupes qui ont une position basale dans la phylogénie des Nématodes (Figure 6). Ils affichent la plus faible diversité d'espèces parasites de plantes. Les NPP des clades 1 et 2 sont des ectoparasites migrateurs vecteurs de virus. Ils infligent de faibles dégâts directement via leur processus d'alimentation mais des dégâts beaucoup plus importants via la transmission de virus.

  On distingue le clade *Triplonchida* (clade 1) qui correspond à des espèces parasites des racines, tubercules et rhizomes du clade *Dorylaimida* (clade 2) correspondant à des parasites des racines uniquement.

- Le clade 10 correspond au clade *Aphelenchida* (clade10 b, Figure 6) qui comprend des espèces parasites des parties aériennes (nématodes foliaires). Ce sont des parasites obligatoires ou facultatifs qui peuvent se nourrir de champignons mais aussi des parties aériennes des plantes. Des insectes sont vecteurs de certains de ces nématodes (par exemple, le nématode du pin *Bursaphelenchus xylophilus*).
- Le clade 12 des *Tylenchida* est le groupe le plus distal dans la phylogénie des Nématodes (Figure 6). Il comporte la plus grande diversité d'espèces de NPP mais aussi les espèces les plus dommageables pour l'agriculture. L'ensemble des espèces de ce clade infectent les parties racinaires de la plante. La particularité de ce clade est qu'il a vu l'émergence au moins cinq fois indépendamment d'un mode de vie endoparasite sédentaire à partir d'ancêtres migrateurs ecto ou semi-endoparasites.

Il est à noter que les clades 10 et 12 comportent les espèces les plus destructrices mais aussi les plus largement distribuées, c'est pourquoi ces clades sont les plus étudiés, les mieux décrits dans la littérature et ceux pour lesquels on dispose du plus grand nombre de données (Holterman et al., 2017).

Les espèces parasites de plantes se trouvent intercalées par de très nombreuses espèces arborant différents modes de vie libres ou parasites d'animaux. Un mode de vie phytophage a donc évolué indépendamment à plusieurs reprises, au minimum 3 (Blaxter et Koutsovoulos, 2015) ou 4 fois (van Megen et al., 2009) selon le consensus des différentes phylogénies. Bien que sensiblement différente, la phylogénie de Van Megen et collaborateurs confirme le caractère polytomique et l'évolution multiple et indépendante vers un mode de vie phytophage (Figure 6). De même, la dernière étude en date (Ahmed et al., 2022) comprend plus de 300 espèces et repose sur l'analyse combinée de caractères morphologiques et moléculaires (super-matrice concaténant plusieurs gènes et incluant les séquences de la petite sous-unité de l'ADN ribosomal 18S ainsi que le génome mitochondrial). Les résultats

obtenus avec ces différents caractères confirment la topologie générale des précédentes phylogénies des nématodes (Figure 6) et le caractère polytomique (Figure 7) des NPP (Ahmed et al., 2022).

Si on analyse plus en détail les 4 clades contenant des NPP, il semble que la capacité à parasiter les plantes soit un caractère qui a été acquis et perdu à de nombreuses reprises dans ce phylum. En effet, on observe au moins cinq transitions entre les modes trophiques fongivores et phytoparasites dans le clade 10 par exemple. Au sein du clade 12, il semble qu'une seule transition ait eu lieu vers un mode de vie parasitaire de plantes. Cependant, on observe une transition inverse au sein de ce clade 12 depuis un mode parasite de plante vers un mode parasite d'insectes dans le sous-ordre monophylétique des *Hexatylina* (Holterman et al., 2017) De plus, il est intéressant de remarquer que, parmi les espèces de ce clade qui sont exclusivement des phytoparasites obligatoires, certaines sont sédentaires mais d'autres sont migratrices, et la sédentarisation résulte de multiples émergences indépendantes. La reconstruction phylogénétique de ce groupe met en évidence de multiples apparitions et pertes du mode de vie sédentaire au sein de ce clade (Ahmed et al., 2022; Holterman et al., 2017).

Ainsi, les principales analyses phylogénétiques du phylum Nematoda publiées jusqu'ici présentent les nématodes parasites de plantes comme un ensemble de taxa polyphylétiques (Figure 5, 6, 7). La capacité à parasiter les plantes a évolué de manière indépendante chez les nématodes en suivant des trajectoires évolutives convergentes.

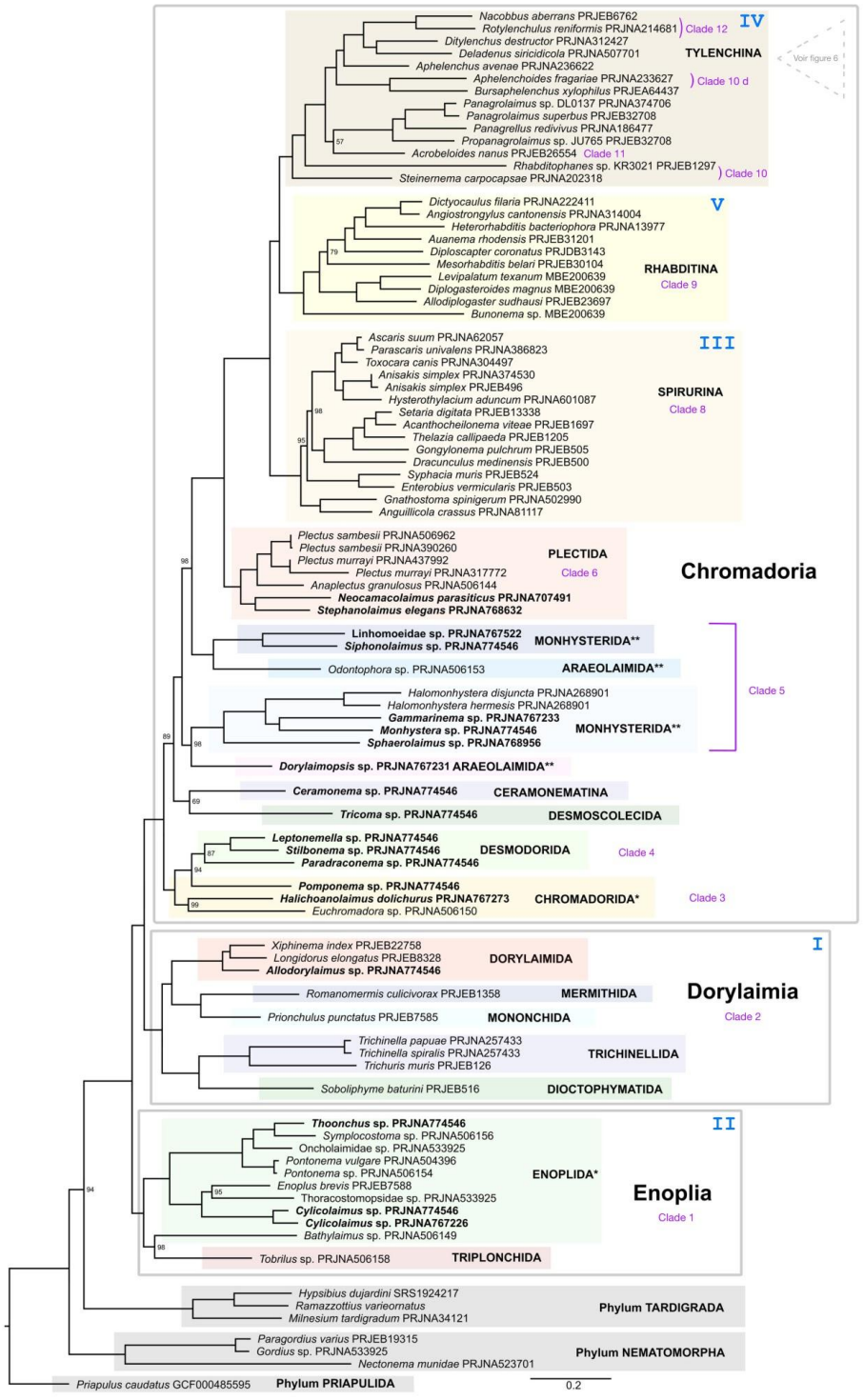


*Figure 6. Phylogénie générale du clade Nematoda inférée par la méthode de maximum de vraisemblance implémentée dans IQ-TREE utilisant le modèle PMSF (LG + C20 + G + F). Seules les valeurs de bootstrap < 100 % sont présentées. Les ordres paraphylétiques sont marqués par *, les ordres polyphylétiques sont marqués par **. Les clades correspondant à la classification de Van Megen sont indiqués en violet et les clades correspondant à la classification de Blaxter sont indiqués avec des chiffres romains de couleur bleu. Figure modifiée issue de (Ahmed et al., 2022).*

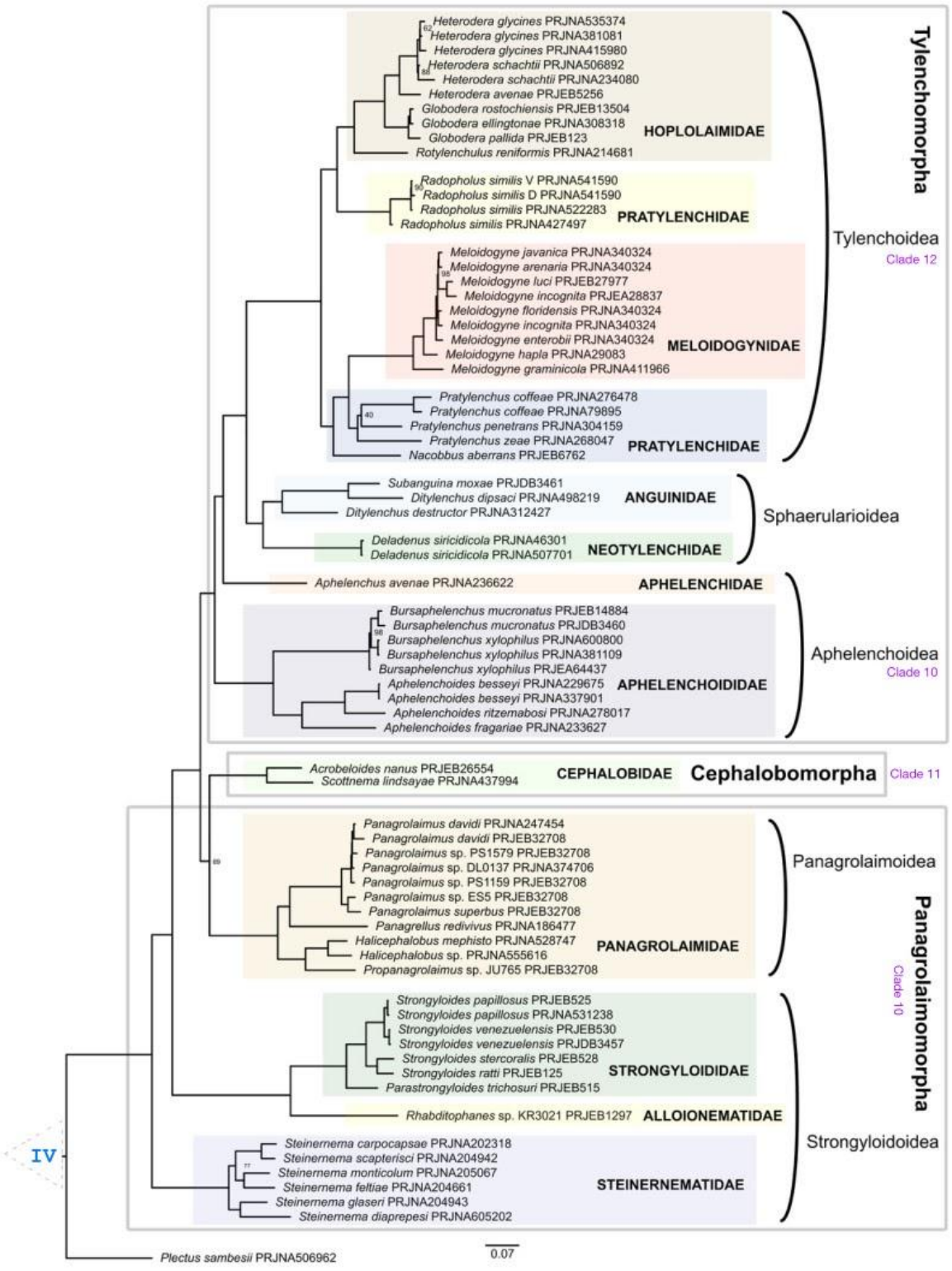


Figure 7. Phylogénie du clade Tylenchina inférée selon la méthode du maximum de vraisemblance implémentée dans IQ-TREE (seules les valeurs de soutien bootstrap < 100% sont indiquées). Les clades correspondant à la classification de Van Megen sont indiqués en violet et les clades correspondant à la classification de Blaxter sont indiqués avec des chiffres romains de couleur bleu. Figure modifiée issue de (Ahmed et al., 2022).

5. *Mécanismes d'infection et cycle de vie*

Les nématodes parasites de plantes présentent une grande diversité morphologique (Figure 8) et une large variété de mécanismes d'interaction avec leurs hôtes (Agrios, 2005). Ils possèdent tous un stylet buccal creux et protubérant en forme d'aiguille permettant au nématode d'injecter des sécrétions salivaires, et d'absorber le contenu des cellules végétales (Baldwin et al., 2004). Ces sécrétions contiennent des protéines et autres molécules appelées effectreurs, permettant de déjouer les systèmes de défense de l'hôte ainsi que de reprogrammer son développement et sa physiologie (Mitchum et al., 2013). On retrouve ce processus parasitaire de sécrétion d'effecteurs aussi chez les champignons (Fei et Liu, 2022) et les oomycètes (Evangelisti et al., 2013). Chez les nématodes, il existe plusieurs types de stratégies parasitaires que l'on peut distinguer selon le processus d'alimentation et le mode de vie.

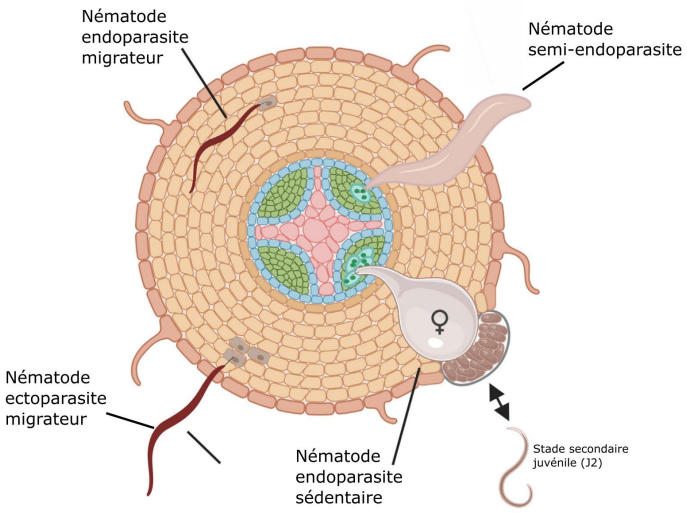


Figure 8. Les différents modes de vie et d'alimentation des nématodes phytoparasites. Figure modifiée issue de (Topalovic, 2020).

a) Les ectoparasites

Certains NPP sont des ectoparasites migrateurs qui ne pénètrent jamais à l'intérieur de l'hôte, ils migrent dans le sol et utilisent les racines comme source de nourriture occasionnelle quand l'opportunité se présente. À ce titre, ils sont comparables aux insectes piqueurs-suceurs tels que les pucerons. Ainsi, les

ectoparasites, comme les nématodes des genres *Tylenchorhynchus, Xiphinema* et *Tylenchus,* sont toujours à l'extérieur des tissus végétaux et du cortex racinaire (Danchin et al., 2017). Ce type de parasite provoque indirectement de gros dégâts agricoles en tant que vecteurs de virus mais les espèces de NPP les plus dommageables pour l'agriculture sont des espèces endoparasites capables de pénétrer à l'intérieur des tissus végétaux (Jones et al., 2013).

b) Les endoparasites migrateurs

Les endoparasites migrateurs pénètrent dans l'hôte et migrent dans les tissus de l'hôte ce qui cause généralement des dommages physiques importants dans la plante, comme c'est le cas notamment pour les nématodes des genres *Radopholus* ou *Pratylenchus* (Jones et al., 2013). En revanche, les nématodes semi-endoparasites peuvent avoir des stades migrateurs mais peuvent aussi se fixer dans la plante hôte afin de se nourrir à un stade de leur cycle de vie. C'est le cas des espèces des genres *Tylenchulus, Helicotylenhcus* et *Scutellonema*. Ces espèces migrent dans les racines mais présentent aussi un stade sédentaire où elles induisent une structure nourricière.

c) Les endoparasites sédentaires

Finalement, les espèces les plus étudiées sont celles qui ont un mode de vie sédentaire car ce sont celles qui présentent le plus fort impact économique. En effet, les espèces sédentaires, comprenant les nématodes à kyste et les nématodes à galles, sont les espèces les plus nuisibles (Jones et al, 2013 ; Barthlem et al.,2014) et présentent les stratégies d'alimentation les plus complexes (Gheysen et Mitchum, 2011). Ces biotrophes induisent le développement de structures d'alimentation complexes dans les racines végétales qui fournissent une source de nourriture stable (Figure 8, 9). Pour ces espèces sédentaires, les œufs sont libérés dans le sol mais les femelles adultes restent au même endroit dans la racine pendant toute leur vie et meurent en relâchant les œufs (Jones et al., 2013).

Les nématodes du genre *Meloidogyne* induisent le développement de galles racinaires (Jones et Payne, 1978), raison pour laquelle ils sont appelés nématodes à galles. Ces galles comprennent des cellules végétales géantes multinucléées provenant de divisions successives des noyaux sans division des cellules elles-mêmes. Ces cellules géantes sont métaboliquement très actives et servent de site nourricier aux nématodes (Figure 9 A,C). D'autre part, les nématodes à kystes des genres *Heterodera* et *Globodera* établissent des syncytia dans les racines des plantes (Jones, 1981). Ici, le site nourricier provient de la fusion de cellules végétales adjacentes en syncytium (Figure 9 B, D) (Szakasits et al., 2009).

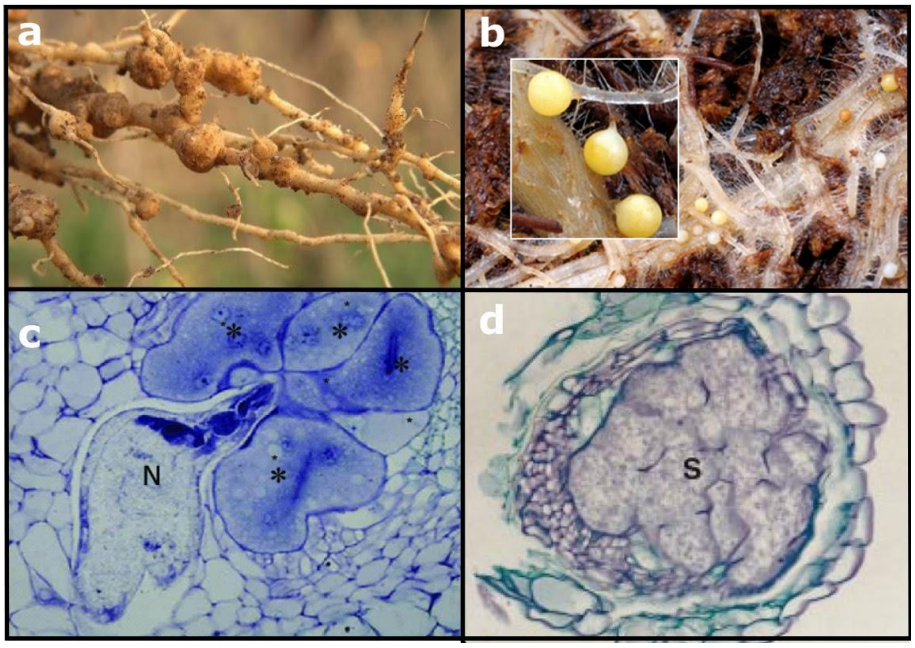


Figure 9. Une infection par des NPP endoparasites induit la formation de structures racinaires visibles correspondant à (A) des galles en cas d'infection par des nématodes du genre Meloidogyne; ou (B) des kystes pour les nématodes des genres Globodera et Hétérodera. Les nématodes induisent ces déformations en modifiant l'activité cellulaire végétale pour former un site nourricier constitué (C) de cellules géantes (les galles) pour les nématodes du genre Meloidogyne, et (D) d'un syncytium cellulaire pour les nématodes à kystes. Photographies issues de (Miyara et al., 2015).

Ainsi, certaines fonctions biologiques comme la capacité de reprogrammer des cellules racinaires pour les transformer en sites métaboliquement hyperactifs qui deviennent alors source de nutriments sont spécifiques à certains clades de

nématodes sédentaires (Siddique et al., 2009, 2012 ; Szakasits et al., 2009 ; Hofmann et al., 2010). D'autres fonctions biologiques comme la dégradation de la paroi des plantes sont communes à l'ensemble des espèces car elles doivent toutes faire face aux mêmes challenges liés au contournement des défenses des plantes et à la métabolisation des produits de la plante.

La phytophagie est un comportement alimentaire hautement sophistiqué qui nécessite des capacités de contournement des défenses des plantes, la détoxification et la digestion des macromolécules végétales (Roberts et al., 1988). L'évolution vers ce mode de vie parasitaire a nécessité l'émergence d'innovations génétiques au sein de ces clades. Les processus évolutifs à l'origine de ces innovations ne sont pas encore totalement compris mais plusieurs mécanismes semblent y avoir contribué.

6. Principaux processus évolutifs impliqués dans le parasitisme des plantes

De manière générale, l'adaptation[1] à un nouveau mode de vie ou à des changements environnementaux repose sur l'apparition d'une nouvelle caractéristique physiologique ou phénotypique. L'apparition de ce trait chez un individu provient principalement des variations de la séquence génomique qui impactent la régulation de l'expression génique ou une fonction protéique. Les mutations génétiques à l'origine de ces variations (substitution, insertion, délétion, inversion…), qu'elles soient ponctuelles ou à plus grande échelle, peuvent toucher toutes les régions du génome. Ces variations génétiques apparaissent d'abord chez un individu et sont ensuite transmises à la descendance puis propagées ou rejetées dans les populations par sélection naturelle. Plusieurs processus évolutifs résultant en une modification de la séquence génétique d'un individu par gain ou perte de matériel génétique et/ou par accumulation de mutations ponctuelles de la séquence

[1] Le terme 'adaptation' désigne ici le processus évolutif aboutissant à l'apparition de nouveaux phénotypes dans une population puis une espèce. Il est à noter que ce processus adaptatif se distingue du processus d'acclimatation qui désigne l'adaptation rapide et réversible d'un individu à des changements environnementaux reposant sur des processus de plasticité phénotypique.

ancestrale ont déjà été décrits dans la littérature. Nous en présentons ici un résumé synthétique.

a) La perte de gènes

La perte de matériel génétique peut induire des variations phénotypiques qui sont généralement neutres ou délétères mais qui, dans certains cas et sous certaines conditions, peuvent résulter en un avantage sélectif. Cette source de variation génétique semble avoir façonné l'évolution de tous les règnes de la vie (Albalat et Cañestro, 2016). Chez de nombreux pathogènes, bactériens et fongiques, la perte de gènes constitue un mécanisme d'adaptation qui intervient de manière récurrente dans les processus évolutifs liés à l'adaptation pour faire face au mécanismes de défense développés par l'hôte (McNally et al., 2016). Il semble que ce type d'événement génomique soit aussi impliqué dans les processus adaptatifs des NPP. En effet, Castagnone-Sereno et collaborateurs ont montré que, chez *Meloidogyne incognita,* la perte de copies de gènes peut être associée à la capacité des nématodes à contourner la résistance de la plante hôte sous certaines conditions (Castagnone-Sereno et al., 2019).

b) L'acquisition de gènes

L'apparition de traits phénotypiques correspondant à une nouvelle fonction biologique repose généralement sur un gain de gènes. Différents mécanismes évolutifs à l'origine de l'émergence de nouveaux gènes ont été décrits dans la littérature. On distingue les processus de *(i)* néofonctionnalisation par duplication et cooptation, de *(ii)* l'émergence de gènes *de novo* à partir de régions non géniques, de *(iii)* l'acquisition de gènes provenant du génome d'un autre individu par transferts horizontaux.

(i) La duplication suivie de néofonctionnalisation

Le processus évolutif de duplication suivie de la néofonctionnalisation d'une des deux copies désigne l'apparition d'un gène possédant une nouvelle fonction biologique à partir d'un gène existant. Suite à l'accumulation de mutations, un gène peut avoir un impact biologique différent par la modification de son expression, de la fonction de la protéine qu'il code, ou encore de la localisation subcellulaire de la protéine codée. Ce processus comprend souvent une étape de duplication du gène suivie d'une phase de diversification/spécialisation par mutation de l'un des gènes paralogues. On parle alors de processus de néofonctionnalisation (True et Carroll, 2002).

De nombreuses études mettent en évidence la contribution des mécanismes de duplication et de néofonctionnalisation dans l'émergence de nouveaux gènes et ainsi dans l'acquisition de fonctions biologiques essentielles aux processus adaptatifs (Levasseur et al., 2007).

Il existe une famille de protéines appelée « antigel » que l'on retrouve chez de nombreuses espèces de bactéries, de champignons, de plantes et d'animaux. Ces protéines se lient aux cristaux de glace et empêchent leur développement ce qui confère à l'organisme une protection contre le gel (Tomimatsu et al., 1976). En 2010, une étude portant sur des poissons de la famille des *Zoarcidae*, communément appelés lycodes ou loquettes, a mis en évidence pour la première fois la néofonctionnalisation d'un gène codant pour la protéine antigel de type III. Dans cette étude les auteurs montrent que le gène ancestral comportait un domaine de liaison à la glace et un domaine acide sialique synthase. Après duplication, l'un des paralogues de ce gène a accumulé des mutations aboutissant à la substitution du domaine acide sialique synthase par un peptide signal permettant une optimisation de la fonction antigel. Ces travaux ont montré que les fonctionnalités annexes d'un gène peuvent, suite à duplication et accumulation de mutations, mener à une nouvelle protéine permettant l'adaptation à un nouvel environnement, et donnent un

aperçu de la façon dont les copies de gènes confrontées à des pressions de sélection et de mutation, peuvent emprunter des voies évolutives divergentes (Deng et al., 2010).

Le processus de néofonctionnalisation a aussi contribué à l'innovation génétique et à l'adaptation de certaines lignées de nématodes parasites de plantes. L'étude de Lilley et collaborateurs a montré qu'au cours de l'évolution des *Tylenchida*, le gène de ménage de la *glutathion synthétase* a été dupliqué et que les mutations accumulées par les copies paralogues ont donné naissance à des effecteurs de type *glutathion synthétase*. En effet, certaines copies de gènes ont acquis un peptide signal pour la sécrétion et plusieurs éléments promoteurs modifiant leur expression spatiale et temporelle par la glande sécrétrice dorsale du nématode. Ainsi, cette étude témoigne de la réaffectation d'un gène de ménage endogène pour former une famille d'effecteurs présentant une activité biochimique très diversifiée qui représente une nouvelle classe d'enzymes impliquées dans le parasitisme (Lilley et al., 2018). Cet exemple de recrutement d'une copie de gène pour des fonctions parasitaires met en évidence la contribution de ce processus évolutif de duplication et néofonctionnalisation dans l'émergence de certaines fonctions parasitaires chez les NPP.

(ii) L'émergence de gènes de novo

Selon la théorie communément admise, la duplication de gènes existants constituait le principal processus d'innovation génétique (Ohno, 2013). Cependant, l'acquisition de données génomiques a fourni de plus en plus d'éléments démontrant des processus d'innovation génétique impliquant des régions précédemment non géniques.

En 2005, le Dr. Mar Albà a montré que les gènes plus récents ont tendance à évoluer plus rapidement. Elle a alors formulé l'hypothèse que ces gènes pourraient avoir émergé *de novo* à partir de régions non géniques, expliquant ainsi la moindre adaptation de ces gènes et le fait qu'ils pourraient nécessiter davantage

d'ajustements. Un gène qui aurait émergé *de novo* apporte une nouvelle fonction sans lien avec la fonction du gène ancestral comme c'est le cas pour les gènes issus d'un processus de duplication (Albà MM, 2005). Les premiers exemples documentés concernent l'émergence de gènes *de novo* impliquée dans la sexualité dans le génome de la drosophile (Begun et al., 2007, 2006). Ces travaux pionniers ont suscité beaucoup de scepticisme au sein de la communauté scientifique mais depuis la littérature à ce sujet n'a cessé de s'enrichir (Van Oss et Carvunis, 2019). Chez de nombreuses espèces eucaryotes, une partie substantielle des génomes est non génique et certaines de ces régions ont une composition moléculaire similaires à celles des gènes (i.e. composition en GC) (Palazzo et Lee, 2015).

Comme mentionné précédemment, chez certaines espèces de poissons de la famille des *Zoarcidae*, l'émergence de protéines « antigel » repose sur un processus de duplication suivie de la néofonctionnalisation d'une des deux copies. En revanche, chez les poissons de la famille des *Gadidae* (ou Cabillaud) les gènes codant pour les glycoprotéines antigel semblent avoir émergé *de novo* (Baalsrud et al., 2018). La description de ce processus évolutif chez le cabillaud a été le premier exemple de gène essentiel né d'un ADN non génique chez une espèce non modèle. Plus récemment, des études ont mis en évidence l'implication de gènes *de novo* dans des phénotypes résistants ou pathogéniques de certains micro-organismes. Par exemple, les souches de papillomavirus impliquées dans le développement de cancers humains comportent spécifiquement un oncogène appelé E5 qui a émergé indépendamment dans le génome de chaque souche virale à partir d'une région non codante (Willemsen et al., 2019). L'étude de différentes variétés de riz a montré que les gènes issus de mécanismes *de novo* peuvent représenter 10% des gènes acquis dans le génome de la variété *Oryza sativa Japonica* (Zhang et al., 2019).

Les innovations génétiques impliquées dans l'évolution vers un mode de vie parasitaire proviennent certainement en partie de l'émergence de gènes de novo. Suite à l'analyse du transcriptome de *P. penetrans*, Vieira et collaborateurs suggèrent

que plusieurs gènes codant pour des effecteurs spécifiques à ce genre de NPP ont évolué *de novo* (Vieira et al., 2020). À ce jour, ce travail est le seul à notre connaissance qui décrit des gènes comme ayant émergé *de novo* dans un génome de NPP. Pourtant, chez les nématodes du genre *Meloidogyne* de nombreux gènes n'ont aucun homologue identifiable ailleurs dans le vivant et pourraient être apparus à partir de région non géniques. En 2013, une étude comparative portant sur les génomes du genre *Meloidogyne* a identifié 15 952 gènes de nématodes conservés dans les génomes d'espèces parasites de plantes et absents du reste du vivant. L'analyse fonctionnelle des protéines codées par ces gènes appelés orphelins a permis d'en identifier douze dont les fonctions sont directement impliquées dans le succès de l'interaction parasitaire avec des plantes (Danchin et al., 2013). Ces résultats sont confirmés par une analyse phylogénomique récente portant sur soixante et un génomes de nématodes dont seize appartenant à des espèces parasites de plantes. Cette dernière analyse a identifié plus de 24 000 familles de protéines spécifiques à ces parasites et n'ayant aucun homologue dans les bases de données (Grynberg et al., 2020). La majorité de ces gènes orphelins sont effectivement supportés par des données transcriptomiques et conservées chez différentes espèces. Ils pourraient avoir émergé *de novo* depuis les régions non géniques du génome ancestral de ces parasites. Il ne faut cependant pas écarter l'hypothèse d'une duplication suivie d'une divergence importante ne permettant pas la détection d'homologie ou encore l'acquisition par transfert horizontal depuis les génomes d'espèces non représentés dans les librairies de séquences. L'étude du mécanisme évolutif à l'origine des nombreux gènes orphelins trouvés chez les NPP nécessite de disposer de génomes de très bonne qualité avec un assemblage contigu pour des espèces proches ne possédant *a priori* pas ces gènes. Cela a pu limiter, jusqu'à aujourd'hui, l'exploration de cette question chez les NPP.

Enfin, le gain de traits par acquisition de gènes peut aussi se faire via le transfert horizontal de matériel génétique entre organismes.

3. *Les transferts horizontaux de gènes*

A. Définition et impact évolutif

1. *Définition*

Le matériel génétique est communément transmis verticalement des individus parentaux à leur descendance selon les voies de l'hérédité (Campbell et al., 2004; Osborn, 1902). Cependant, les études de génomique comparative ont rapidement révélé l'existence de transfert horizontaux de gènes (THG) dans lequel du matériel génétique est transféré entre individus de même génération selon des mécanismes indépendants de l'hérédité (Figure 10).

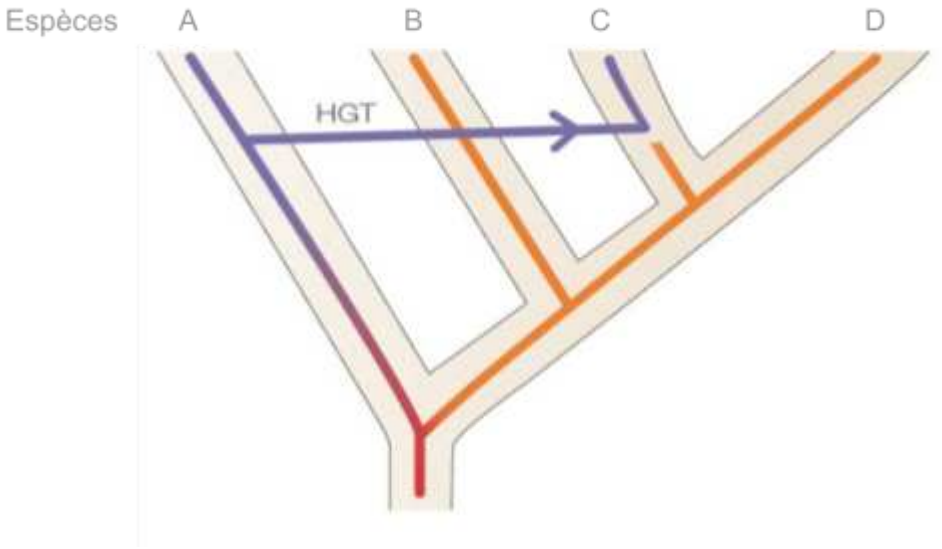


Figure 10. Schéma de TH entre individus ancestraux de A vers C. Ce processus de TH peut avoir lieu entre individus de même espèce (TH intraspécifique), mais il peut aussi concerner des individus de taxa plus ou moins éloignés pouvant même appartenir à des règnes taxonomiques différents (TH interspécifique).

Lorsque les fragments d'ADN transférés contiennent des gènes, ces derniers peuvent, dans certains cas, conserver leur fonctionnalité et fournir de nouvelles aptitudes permettant l'adaptation à un changement environnemental ou à un autre mode de vie. Bien que la machinerie transcriptionnelle soit différente entre organismes eucaryotes et procaryotes, le THG peut permettre l'acquisition de nouvelles fonctions grâce à l'universalité[2] du code génétique. Si le THG est viable, il

[2] À l'exception de certaines espèces de virus, de phage, de bactéries et d'eucaryotes unicellulaires.

pourra être transmis aux générations suivantes et se fixer dans la population par dérive génétique ou sélection naturelle s'il fournit un avantage sélectif au receveur.

Au sein d'un même génome, les gènes peuvent avoir des histoires évolutives différentes, en particulier en cas de THG. Ces trajectoires évolutives ne peuvent pas être simplement représentées sous forme d'un arbre phylogénétique. C'est pourquoi certains auteurs proposent une représentation sous forme de réseaux plutôt qu'un arbre de la vie afin de tenir compte de ce type d'événements et de reconstruire plus précisément les liens de parentés entre individus (Husnik et McCutcheon, 2018; Syvanen, 1987).

Le THG est pleinement accepté comme mécanisme d'évolution depuis de nombreuses années chez les procaryotes (Jain et al., 1999, Lawrence, 1999, Ochman et al., 2000). En revanche, ce mécanisme a longtemps été controversé lorsqu'il implique des organismes eucaryotes qui possèdent un noyau et, dans le cas des animaux, lorsque la lignée somatique et germinale sont séparées (Martin, 2017).

2. *Exemple d'impact évolutif, cas des protéines antigels*

Comme nous l'avons vu précédemment, les THG semblent avoir également contribué à l'acquisition et la diffusion de gènes codant pour les protéines « antigel » chez les poissons à nageoires rayonnées et à leur adaptation à des conditions environnementales extrêmes. Dans le clade des *Actinoptérygiens*, cinq types de protéines « antigel » très différentes ont été identifiées avec une distribution éparse chez des espèces très éloignées, comme par exemple les protéines antigel de type II retrouvées chez le Hareng, Éperlan et Corbeau de mer. Selon les chercheurs travaillant sur l'évolution de ces espèces, la phylogénie de ces protéines est un véritable casse-tête car :

- Dans les génomes de lignées de poissons très proches d'un point de vue évolutif, on retrouve des séquences codant pour une protéine antigel très divergente.

- Dans les génomes de poissons très éloignés phylogénétiquement, on retrouve des séquences codant pour une protéine antigel très similaires.

En raison de l'absence de corrélation entre la distance évolutive et le niveau divergence nucléotidique entre les gènes, les auteurs ont formulé l'hypothèse que ce gène a transité horizontalement entre les génomes de ces poissons (Graham et al., 2012). Suite à l'assemblage récent du génome du hareng, l'analyse de l'environnement génomique des gènes codant pour la protéine antigel conforte cette hypothèse et suggère qu'une copie de ce gène a été transférée horizontalement du hareng à l'éperlan (Graham et al., 2012; Graham et Davies, 2021).

L'exemple de la protéine antigel chez les poissons illustre comment différents mécanismes évolutifs contribuent de manière conjointe à l'évolution des espèces, à leur adaptation à des changements environnementaux ou à une nouvelle niche écologique. Différents processus évolutifs ont probablement contribué conjointement à l'innovation génétique dans les génomes de nématodes et l'émergence de leur capacité à parasiter les plantes car on retrouve aussi chez ces espèces des gènes acquis par transfert horizontal.

B. L'étendue du processus de THG dans le vivant

1. *Découverte des THG chez les organismes procaryotes*

a) L'expérience de Griffith, observation d'un processus de transformation bactérienne

Le transfert horizontal de gènes a été découvert chez des bactéries puis plus largement chez les procaryotes et d'autres organismes unicellulaires. Ce processus a été mentionné pour la première fois en 1928 dans l'expérience historique de Griffith qui consistait à infecter des souris avec différentes souches (virulente/avirulente - active/inactive) de bactéries de l'espèce *Streptococcus pneumoniae*.

Lors de cette expérience, Griffith observa que la contamination de souris par une souche avirulente ayant été en contact avec une souche virulente inactivée, provoque la mort des souris infectées. Sur la base de ces observations, Griffith a émis l'hypothèse qu'un composant chimique des cellules virulentes avait en quelque sorte transformé les cellules non virulentes en la forme la plus virulente (Griffith, 1928).

À cette époque, les connaissances en biologie moléculaire étaient insuffisantes pour comprendre précisément ce mécanisme de transformation naturelle qui sera révélé plus tardivement.

b) L'ADN comme support de l'information génétique et compréhension du processus de THG

Ce n'est qu'en 1944 que Avery et collaborateurs reproduisent l'expérience de Griffith en montrant que le vecteur transformant est une molécule d'ADN. Ces résultats obtenus grâce aux progrès de la biochimie suggèrent que l'ADN est le support de l'information génétique. Les auteurs font alors l'hypothèse que le transfert de fonction biologique observé a eu lieu par des mécanismes indépendants de l'hérédité (Alačević, 1963; Kasuya, 1964; Sermonti et Spada-Sermonti, 1955).

Ces travaux contestés seront plus largement acceptés suite aux travaux pionniers de Rosalind Franklin et à la découverte de la structure tridimensionnelle de l'ADN par James Watson et Francis Crick. Grâce à ces découvertes et aux progrès en génétique, l'impact des THG dans l'histoire évolutive des espèces a été plus largement considéré par la communauté scientifique.

Depuis ces découvertes, la liste de cas de THG recensés impliquant des organismes Procaryotes n'a cessé de s'allonger et des THG impliquant des *Archées* ont aussi été rapportés (Wagner et al., 2017). Les mécanismes impliqués dans les transferts de matériel génétique chez les procaryotes sont aujourd'hui bien décrits (conjugaison, transformation, transduction…). Et on sait que, chez ces organismes, les événements sont très fréquents et ont des impacts évolutifs très vastes.

2. *Fréquence des THG chez les organismes unicellulaires*

Chez les organismes procaryotes et donc dépourvus de noyau, le matériel génétique flotte dans le cytoplasme et semble plus facilement accessible que chez les organismes eucaryotes. Si l'hypothèse de THG a rapidement été acceptée chez les procaryotes, ce processus évolutif a été plus longuement discuté en ce qui concerne la lignée eucaryote et certains scientifiques restent encore aujourd'hui sceptiques à ce sujet (Martin, 2017). Cependant, de plus en plus d'indices s'accumulent en faveur de la théorie d'un flux de gènes pouvant impliquer toutes les espèces. A l'avenir, l'émergence des technologies de séquençage longues lectures pourra confirmer l'intégration au sein des génomes des THG présumés.

Il semble même que ce processus de THG ait joué un rôle majeur dans l'émergence des lignées eucaryotes actuelles. Comme mentionné dans la partie « Contexte général », selon la théorie de l'endosymbiose les cellules eucaryotes sont nées de l'incorporation de bactéries (protéobactérie: mitochondrie chez les animaux, cyanobactérie: chloroplaste chez les plantes) dans une cellule ancestrale appartenant au clade des Archées (Eme et al., 2017). Afin de permettre une relation durable entre les partenaires, plusieurs études ont montré que des THG avaient eu lieu du génome de la mitochondrie vers le génome nucléaire de la cellule hôte ancestrale (Ponce-Toledo et al., 2019; Wei et al., 2022). Le même processus est observé concernant l'endosymbiose secondaire avec un flux de gène des plastes vers le génome nucléaire (Filip et Skuza, 2021). Les THG ont donc joué un rôle crucial dans la transition entre endosymbionte et organelle. Ce même phénomène de THG est aussi observé entre les génomes des parasites intracellulaires et leurs cellules eucaryotes hôtes, comme c'est le cas par exemple chez les parasites du genre Microsporidia et leurs cellules hôtes animales (Corradi, 2015).

L'accumulation de données génomiques fournit un meilleur aperçu de la distribution des gènes aux seins des génomes des espèces vivantes et permet ainsi de mieux reconstruire leur histoire évolutive suggérant que de nombreux événements

de THG ont ponctué l'évolution du vivant. Par exemple, on sait que des THG ont eu lieu entre mitochondrie et génome nucléaire juste après l'intégration cellulaire mais une étude récente démontre que certains événements de transferts sont plus récents. En effet, Wei et collaborateurs ont analysé les génomes de plus de 60 000 individus à la recherche d'événements d'insertion de gènes mitochondriaux. Ce travail a révélé des insertions de fragments de génomes mitochondriaux dans le génome nucléaire de l'homme, dont la majorité des événements de transferts ont eu lieu après le phénomène de spéciation entre l'homme et les primates. Ces observations indiquent que ce flux de gènes entre mitochondrie et génome nucléaire est un processus continu. Aussi, cela montre que des fragments d'ADN étrangers peuvent pénétrer les noyaux des cellules eucaryotes plus fréquemment qu'on ne l'imaginait jusqu'à aujourd'hui (Wei et al., 2022). De même, il semble que des processus de TH de matériel génétique peuvent avoir lieu de manière récurrente entre cellules d'un même organisme. Par exemple, les cellules présentatrices d'antigènes sont capables de couper leurs télomères et de les 'donner' aux cellules T via des vésicules pour ralentir le processus de sénescence et prolonger l'immunité. De cette manière, les télomères peuvent être rallongés jusqu'à 30 fois plus qu'avec la seule action de la télomérase (Lanna et al., 2022). Ainsi, les échanges de matériel génétique entre cellules eucaryotes suivies d'intégration dans le génome de la cellule receveuse sont beaucoup plus fréquents qu'on ne l'imaginait il y a quelques décennies et même quelques années.

Les THG sont très fréquents chez les organismes unicellulaires en général et ce processus contribue largement aux processus évolutifs (Sibbald et al., 2020). Comme les organismes procaryotes, une majorité d'organismes unicellulaires eucaryotes ont une reproduction asexuée.

La variabilité génétique dans ces populations repose principalement sur le mécanisme de mutation génétique et, sans brassage génétique, la dissémination de l'information est limitée. Les mécanismes de THG permettent de pallier cette

limitation et de propager une variation génétique avantageuse. Ainsi, les THG sont décrits comme l'un des principaux mécanismes de dissémination de l'information génétique chez ces organismes (Douanne et al., 2022).

Le THG permet aussi la propagation d'innovation génétique apparue chez un individu dans le reste d'une population. Mais il peut aussi permettre d'acquérir directement une fonction existant chez une autre espèce. Chez une espèce de *Saccharomyces* oenologiques, les analyses phylogénétiques ont révélé, par exemple, trois régions génomiques acquises horizontalement d'espèces de levures plus éloignées. Ces régions comprennent 34 gènes qui sont impliqués dans des fonctions essentielles à la fermentation du vin (Novo et al., 2009).

De manière encore plus surprenante, le processus de THG peut aussi permettre à ces organismes de récupérer du matériel génétique provenant de clades beaucoup plus éloignés comme des organismes appartenant à d'autres règnes. L'acquisition de nouveaux gènes par THG peut fournir à certaines lignées eucaryotes des gènes impliqués dans des fonctions biologiques déjà présentes chez d'autres espèces. Par exemple, en 2005, une étude a décrit la présence de gènes bactériens dans le génome d'une espèce d'amibe tellurique (*Dictyostelium discoideum*) se nourrissant de bactéries et de levures. Les gènes acquis par THG permettent à son hôte de résister aux toxines bactériennes (Eichinger et al., 2005).

Le transfert d'ADN dans le génome d'un organisme phagocytaire depuis le génome de sa proie semble être une interaction propice à ce type d'événement (Doolittle, 1998). Cependant, la présence de gènes bactériens dans les organismes non phagocytaires suggère qu'il existe d'autres mécanismes de transfert d'ADN. Quoi qu'il en soit, ces événements de THG peuvent ainsi permettre au receveur eucaryote de se protéger d'autres organismes, de survivre dans de nouveaux environnements ou d'utiliser de nouvelles sources de nourriture.

Une revue de Husnik et McCutcheon montre l'étendue et la récurrence de ce phénomène qui touche la plupart des lignées eucaryotes à de multiples reprises

(Husnik et McCutcheon, 2018). Certaines espèces de champignons, par exemple, ont acquis par THG des gènes impliqués dans de nombreuses fonctions biologiques comme la dégradation des glucides, le métabolisme de l'azote, la survie dans des conditions extrêmes ou le maintien de relations d'endosymbiose (Figure 11).

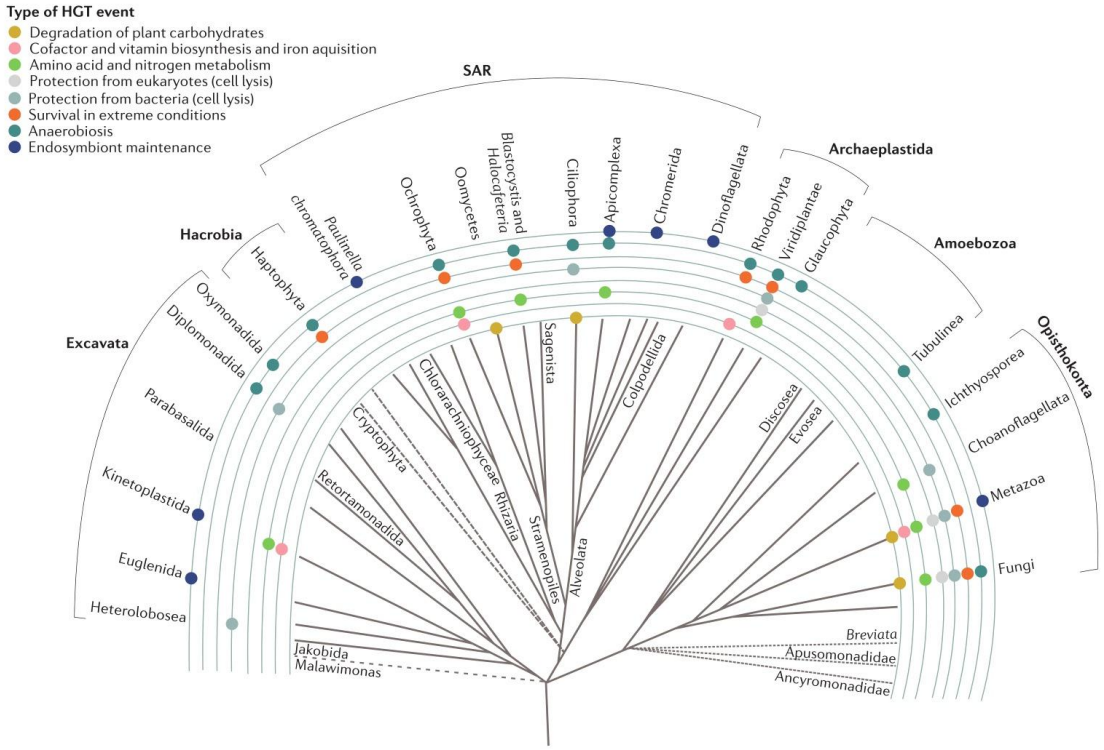


Figure 11. L'arbre phylogénétique des principales lignées eucaryotes indiquant les fonctions biologiques issues de THG à l'aide de cercles de couleurs (à l'exception des gènes provenant des mitochondries et des plastes). Les supergroupes taxonomiques sont indiqués en gras.
Figure issue de (Husnik et McCutcheon, 2018).

Ainsi, les THG semblent avoir eu un impact significatif sur les capacités métaboliques ou l'adaptation des organismes à de nouvelles niches écologiques (Fitzpatrick, 2012). Comme le montre la revue de Husnik et McCutcheon, de nombreux gènes impliqués dans des fonctions biologiques clés ont été acquis par THG chez les organismes unicellulaires mais aussi chez les organismes multicellulaires (i.e. Metazoa, Viridiplantae) (Figure 11).

3. *Les THG chez les organismes multicellulaires*

L'hypothèse de transfert de matériel génétique naturel entre espèces impliquant des organismes eucaryotes multicellulaires a été formulée pour la première fois par Norman Anderson, en 1970. Selon cette hypothèse, les virus peuvent transporter des gènes à travers tous les règnes animaux et végétaux, gènes pouvant être incorporés à l'ADN de l'hôte suivant (Anderson, 1970). Ce n'est cependant que dans les années 80 que les premiers exemples de THG chez les animaux et les plantes ont été rapportés (Syvanen, 1985). Face à la multiplication des cas de THG décrits chez les eucaryotes, Syvanen formule la théorie de transfert de gènes entre espèces la première fois dans la littérature comme phénomène qui « drive » l'évolution génétique (Syvanen, 1985). Dans les années 2000, le nombre de cas décrits va exploser en raison de l'accès aux données génomiques.

L'exploration des génomes et les analyses de génomique comparative ont révélé des THG dans diverses lignées animales et végétales avec une large palette de combinaisons donneurs/receveurs. Les événements de THG semblent avoir ponctué l'histoire de la plupart des espèces mais cela avec des fréquences variables selon les clades. Le flux de gènes impliquant des espèces microbiennes est plus important. Cependant, ce processus peut aussi intervenir entre deux espèces multicellulaires.

a) Les THG impliquant des micro-organismes, des événements plus fréquents

La majorité des THG concernant des génomes d'animaux ou de plantes décrits dans la littérature proviennent de micro-organismes identifiés comme Bactéries (Danchin et al., 2016), Champignons (Kikuchi et al., 2004; Moran et Jarvik, 2010) ou Virus (Gilbert et Belliardo, 2022; Pastuzyn et al., 2018; Pienaar et al., 2022). Les flux de gènes semblent plus fréquents entre les organismes qui entretiennent des relations symbiotiques ou parasites. Un des premiers éléments de compréhension concernant ces observations est que plus le degré de proximité écologique entre les espèces est élevé et les interactions sont étroites, plus la probabilité qu'un THG

survienne est élevée. La majorité des THG décrits dans les génomes d'animaux ou de plantes proviennent de micro-organismes, cependant des transferts peuvent aussi se dérouler entre organismes multicellulaires.

b) THG entre organismes multicellulaires

(i) Transferts entre plantes

Des cas de THG ont déjà été rapportés entre (i) plantes comme, par exemple, les photorécepteurs permettant aux fougères de croître en direction de la lumière et codés par un gène qui n'est retrouvé que chez des embryophytes tel que le Cornifle immergé. L'occurrence de ce gène uniquement dans des clades éloignées suggère que ce gène a transité horizontalement entre des ancêtres de ces lignées végétales (Li et al., 2014).

(ii) Transferts entre animaux

De même, des THG ont également été rapportés chez des lignées animales tel que le cas de la protéine antigel qui semble avoir été transférée entre différentes espèces de poissons comme mentionnés précédemment [partie 2.6.a.a] (Graham et al., 2012). Les transferts de matériel génétique chez ces organismes semblent particulièrement fréquents, vraisemblablement en raison du transit du matériel génétique (gamètes mâles) dans le milieu (pouvant aussi résulter en des phénomènes d'hybridation). Cependant, ce phénomène a aussi été identifié chez des espèces terrestres comme l'illustre par exemple le flux de gènes codant pour le rétrotransposon Bovine-B. En effet, ce gène est habituellement présent chez les mammifères mais il a aussi été retrouvé sporadiquement chez certains amphibiens et reptiles. Une récente étude sur la distribution de ce THG chez ces derniers montre qu'il semble avoir voyagé de la grenouille vers une espèce de serpent prédatrice. L'hypothèse est que ce gène aurait transité entre ces espèces via des parasites (Kambayashi et al., 2022). Ces événements semblent relativement fréquents chez les arthropodes ou les poissons alors qu'ils sont plus rares chez les mammifères (Zhang et al., 2020).

(iii) Transferts entre plantes et animaux

Les gènes peuvent voyager entre espèces animales et végétales mais des travaux rapportent aussi des cas des transferts de gènes depuis les génomes de plantes vers les génomes de certaines espèces animales. Lapadula et collaborateurs ont montré la présence de gènes codant pour des protéines inactivatrices de ribosomes, généralement retrouvés dans les génomes de bactéries ou de plantes, dans le génome de l'insecte ravageur de culture *Bemisia tabaci* (la mouche blanche). Les analyses phylogénétiques soutiennent l'hypothèse que ces gènes ont été acquis par THG d'origine végétale. Les auteurs ont montré que ces gènes, présents dans plusieurs version du génome, présentent des signes de domestications (i.e. transcription et présence d'introns) excluant l'hypothèse d'une contamination (Lapadula et al., 2020). Il est intéressant de noter que des THG codant pour des protéines avec la même fonction biochimique ont aussi été identifiés dans le génome de moustiques de la famille des *Culicidae*, cependant dans cet autre clade d'insectes ces gènes semblent plutôt d'origine bactérienne (Lapadula et al., 2013).

Considérant que seuls les gènes gardés au cours de l'évolution sont observables, le matériel génétique semble transiter entre tous les clades et ces quelques exemples illustrent l'étendue de ce phénomène dans le vivant, ainsi que son impact sur l'évolution des espèces. Les mécanismes en jeu sont encore souvent mal compris en ce qui concernent les transferts impliquant des organismes multicellulaires. La multiplication des travaux portant sur ces questions permet de mieux comprendre ces processus.

C. Les mécanismes de transfert

Les THG impliquent la présence de fragments d'ADN mobiles, leur transit entre donneur et receveur et l'insertion de ces fragments dans le génome de l'hôte.

Pour impacter l'évolution d'une espèce, le matériel transféré devra aussi être transmis aux générations suivantes.

1. *La formation de fragments d'ADN mobiles*

La formation d'un fragment d'ADN mobile et l'insertion dans le génome hôte dépendent de la dynamique des génomes. Chez les micro-organismes, le flux de gènes est facilité par la formation de nombreux fragments d'ADN extra-chromosomiques tels que les plasmides ou les épisomes. Certains de ces fragments, appelés plasmides, peuvent se répliquer de manière autonome. D'autres possèdent des gènes codant, entre autres, pour la synthèse d'enzymes de restriction qui permettent l'intégration au sein des chromosomes hôtes par recombinaison, on parle alors d'épisome.

Ce type de molécule d'ADN extra-chromosomique a été identifié chez les bactéries (Watanabe, 1963), les Archées (Zillig et al., 1996) et eucaryotes unicellulaires (Douanne et al., 2022; Downing et al., 2011; Murray, 1987). Chez les eucaryotes multicellulaires, on distingue aussi différents types d'ADN extra-chromosomiques: l'ADN des organelles et les intermédiaires et/ou sous-produits des transpositions et réarrangements de l'ADN (Rush et Misra, 1985). Par ailleurs, les éléments transposables (ET) sont des séquences d'ADN capables de se déplacer et de se multiplier dans les génomes. Des ET ont été trouvés dans pratiquement toutes les espèces eucaryotes étudiées et peuvent représenter une part importante du matériel génétique (*i.e.* génome humain: ~ 50% ; maïs : 85%) (Boissinot, 2019). L'activité des ET génère des fragments d'ADN extra chromosomiques, ces éléments mobiles sont susceptibles d'intégrer des gènes de leur hôte et d'être véhiculés entre organismes. D'autres mécanismes tels que la recombinaison en présence de séquences répétées peuvent aboutir à la formation d'éléments mobiles dans les cellules eucaryotes (Lanna et al., 2022).

2. *Mécanisme de transport entre cellules*

Une autre étape importante dans le THG, source de nombreuses interrogations, est le transit de l'ADN entre les cellules. Plusieurs mécanismes ont été mis en évidence chez les organismes procaryotes mais en ce qui concerne les organismes eucaryotes notre compréhension reste limitée même si plusieurs éléments démontrent que ce type de transfert est possible.

a) Transduction et transport via un vecteur

Lorsque Norman Anderson a formulé pour la première fois l'idée d'un échange de matériel impliquant des organismes multicellulaires, son hypothèse était que ce matériel pouvait être véhiculé par les particules virales (Figure 10). Ce processus appelé transduction a été mis en évidence pour la première fois par Zinder et collaborateurs avec le système bactéries du genre Salmonella et bactériophage lambda (Zinder et Lederberg, 1952).

Lors de leur cycle cellulaire, les virus peuvent intégrer accidentellement du matériel génétique appartenant à la cellule hôte. Cela a été montré par Loiseau et collaborateurs (Loiseau et al., 2021) en effectuant un suivi du contenu génétique viral pour 11 systèmes hôte-virus différents révélant, de cette manière, l'insertion de matériel génétique de l'hôte dans les génomes viraux (i.e. il s'agissait ici d'ET). Également, Cantoni et collaborateurs ont démontré la formation spontanée de molécules hybrides virus-hôte sous forme de mini cercles lors de l'infection de plants de betterave commune par un virus de la famille *Geminiviridae*. Dans cette étude, les auteurs ont également observé l'encapsidation de ces molécules d'ADN hybride et la dissémination vers les autres cellules de la plante (Catoni et al., 2018). De la même manière, les virus peuvent être vecteurs de matériel génétique entre différents organismes appartenant à leur gamme d'hôte.

On peut alors faire l'hypothèse que, potentiellement, n'importe quel pathogène ou parasite est susceptible de véhiculer du matériel génétique à travers les espèces appartenant à sa gamme d'hôte comme le travail de (Kambayashi et al.,

2022) le suggère. Dans cette étude, les auteurs retrouvent dans les génomes de parasites de reptiles et d'amphibiens des séquences homologues du gène codant pour le rétrotransposon Bovine-B ayant voyagé entre grenouilles et serpents.

b) Conjugaison

Un autre système permettant de transférer du matériel entre cellules, très bien décrit dans la littérature, est la conjugaison. Chez les procaryotes, il existe aussi des mécanismes de transport d'ADN basés sur l'activité d'un complexe protéique appelé système de sécrétion type IV (T4ss). Ce système permet le transit de macromolécules entre le cytoplasme bactérien et le milieu extérieur. Il existe plusieurs familles de T4ss qui ont des mécanismes variables mais que l'on peut regrouper selon leurs différentes fonctions.

L'une des fonctions du T4ss est la conjugaison qui est le processus de THG le plus répandu chez les bactéries. Dans ce mécanisme qui nécessite un contact entre la cellule donneuse et receveuse (Figure 12), le transfert se fait via un système de sécrétion type IV reliant deux cellules bactériennes (Lederberg et Tatum, 1946).

Ce mécanisme de conjugaison semble spécifique aux procaryotes et n'a jamais été décrit dans d'autres clades. Cependant, un mécanisme identique utilisant le système de sécrétion type IV pour le transfert de matériel génétique a été démontré entre la cellule bactérienne et une cellule Eucaryote. Chez la famille de bactéries pathogènes de plante des *Rhizobiaceae*, le système de sécrétion type IV permet à ces bactéries de transférer du matériel génétique bactérien à des cellules végétales (Bitto et al., 2017; Krenek et al., 2015). Actuellement, un tel processus de transfert de matériel génétique entre une cellule procaryote et une cellule animale n'a jamais été observé.

c) Transformation

Ce système de sécrétion type IV joue aussi un rôle central dans le transfert de matériel génétique entre bactéries par conjugaison mais il peut aussi être impliqué

dans le mécanisme de transformation où la bactérie incorpore du matériel génétique présent dans le milieu (Ellison et al., 2018; Furuya et Lowy, 2006; Thomas et Nielsen, 2005).

On parle généralement de transformation pour désigner le mécanisme actif d'absorption d'ADN libre dans l'environnement (Figure 12). Il s'agit du mécanisme qui a été observé chez les bactéries par Griffith 1928 (Ellison et al., 2018; Furuya et Lowy, 2006; Griffith, 1928; Thomas et Nielsen, 2005). Pour que ce processus ait lieu, il faut que les cellules soient compétentes, et cet état est généralement observé suite à un stress environnemental.

Aucun mécanisme de ce type n'a été décrit chez les organismes eucaryotes même si des méthodes de biologie moléculaire ont été développées pour faciliter la pénétration de matériel génétique dans une cellule eucaryote en utilisant, par exemple, du phosphate de calcium. On parle alors plutôt de transfection.

d) Vésiduction

Un quatrième mécanisme de TH, pouvant impliquer à la fois organismes procaryotes et eucaryotes, est récemment venu compléter ces 3 mécanismes historiquement. Ce mécanisme a été nommé **vésiduction** car il implique des vésicules extracellulaires (VE) pouvant transporter du matériel génétique et fusionner avec les membranes d'une cellule étrangère (Figure 12) (Soler et Forterre, 2020). Des VE contenant diverses macromolécules sont générées par tous les organismes vivants et permettent l'échange et la communication intercellulaire entre individus de même espèce ou non.

Douanne et collaborateurs ont montré que les souches résistantes du parasite *Leishmania* produisent des vésicules extracellulaires enrichies en régions génomiques contenant des gènes de résistance relachées dans le milieu. Lors de ce travail, les auteurs ont aussi généré des amplicons circulaires portant des gènes de résistance,

qu'ils ont diffusés dans un milieu contenant des parasites sensibles via des vésicules extracellulaires. Les gènes de résistance se sont propagés chez les parasites sensibles, ce qui a conduit à l'émergence de sous-populations résistantes aux antibiotiques (Douanne et al., 2022). Ces résultats confirment que la théorie de la vésiduction, proposée par Soler et Forterre en 2019, est probablement un mécanisme de THG non seulement chez les procaryotes mais aussi chez les eucaryotes.

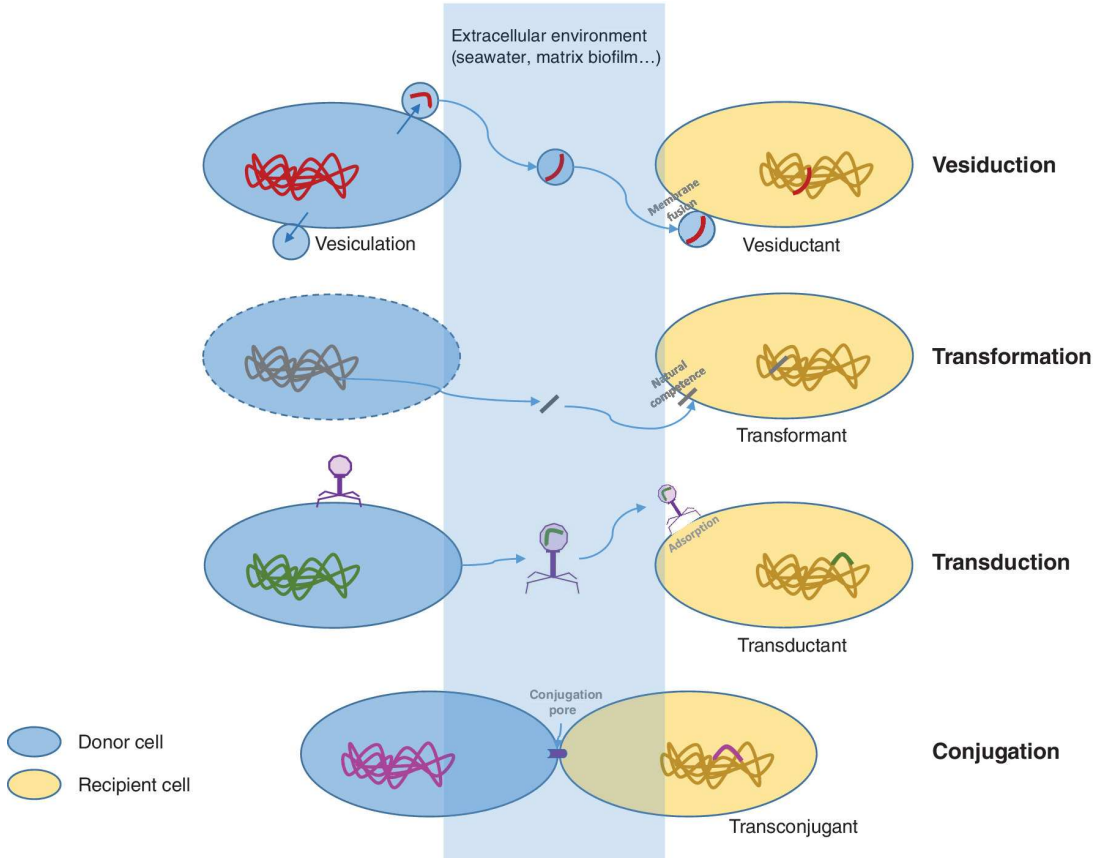


Figure 12. Les mécanismes de transfert de matériel génétique décrits sont : La vésiduction qui implique le transport d'un fragment d'ADN via des vésicules extracellulaires, la transformation d'une cellule compétente qui incorpore un fragment d'ADN présent dans son environnement, la transduction qui se déroule par transport via un vecteur viral, et la conjugaison qui nécessite un contact entre les cellules donneuses et receveuses et l'intervention d'un pore cellulaire procaryote permettant la communication entre les cytoplasmes. Figure issue de (Soler et Forterre, 2020).

On peut aussi noter que Atayade et collaborateurs avaient montré que chez les parasites du genre *Leishmania* ces vésicules extracellulaires peuvent transporter

des endovirus (Atayade, 2019). Sur la base de ces observations, Marcilla et collaborateurs suggèrent que les endovirus peuvent faciliter le transfert et l'insertion de matériel génétique entre cellules proposant un mécanisme appelé Trans-EVduction. Ces études portent sur des organismes unicellulaires mais ce processus peut aussi transférer de l'ADN dans des cellules animales ou végétales (Marcilla, 2022).

e) Autres mécanismes

D'autres hypothèses ont été avancées concernant les différents mécanismes qui peuvent véhiculer de l'ADN entre différents organismes dont potentiellement des animaux. On peut notamment citer l'endosymbiose, la digestion et l'incorporation d'ADN environnemental. Ces hypothèses sont encore discutées bien que de nombreux éléments soient en faveur de l'existence de ces processus. En effet, il y a énormément d'études qui montrent l'échange de matériel génétique entre des endosymbiotes et leurs hôtes comme lors de l'endosymbiose à l'origine de la lignée *Eucaryote* (Ponce-Toledo et al., 2019).

De même pour l'alimentation, l'un des exemples est le cas des limaces de mer de l'espèce *Elysia chlorotica*. Plusieurs études ont suggéré que ces limaces peuvent acquérir des gènes impliqués dans la photosynthèse provenant du génome de l'algue *Vaucheria litorea* dont elles se nourrissent. Cette espèce pratiquant la kleptoplastie est capable de retenir les chloroplastes pendant plusieurs mois dans les cellules de son tube digestif après s'être nourrie de l'algue *Vaucheria litorea* (Green et al., 2000; Trench, 1975). Le maintien de cette organelle nécessite la présence de protéines codées par le génome de l'algue. L'une des théories pour expliquer ce phénomène est que ces gènes sont transférés au génome de la limace. Plusieurs études ont apporté des éléments soutenant cette hypothèse tels que des données biochimiques (Pierce et al., 1996; Rumpho et al., 2008), PCR (Rumpho et al., 2008) et plus récemment par l'analyse transcriptomique (Pierce et al., 2012). Cependant, ces hypothèses ont, ensuite, été réfutées par Bhattacharya et collaborateurs, par l'analyse de génomes de

E. chlorotica ne permettant pas de mettre en évidence l'intégration des gènes en question (Bhattacharya et al., 2013). Les auteurs de ce travail suggèrent que les gènes impliqués dans ce phénomène sont maintenus sous forme de fragments extra-chromosomiques. En 2014, Schwartz et collaborateurs ont publié dans le journal *le bulletin de biologie* les résultats d'une analyse de FISH réalisée avec des sondes correspondant à des gènes de l'algue *V. litorea*. Sur les images, on observe une hybridation des sondes avec des régions chromosomiques mais aussi des régions extrachromosomiques (Schwartz et al., 2014). L'ensemble de ces résultats confirment que des gènes de *V. litorea* subsistent dans les cellules intestinales de *E. Chlorotica* sous forme de fragments extrachromosomiques, au moins partiellement. Cependant, les éléments supportant l'hypothèse d'une intégration de ces fragments au génome de la limace *E. Chlorotica* sont moins évidents.

Les avancées dans le domaine de la génomique environnementale ont montré que du matériel génétique peut transiter dans le milieu comme Graham et collaborateurs le suggèrent dans leur travail sur les THG codant pour la protéine antigel chez les poissons (Graham et Davies, 2021). Certes, ce processus n'a pas été démontré expérimentalement mais en ce qui concerne les espèces animales dont les gamètes ou les œufs sont relargués dans le milieu, l'intégration de matériel génétique à un stade précoce favoriserait par la suite la transmission héréditaire.

3. Transport nucléaire et intégration au génome

Lorsqu'il est question de THG impliquant des cellules eucaryotes, une autre étape qui pose question est l'acheminement de matériel génétique jusqu'au noyau de l'hôte (Martin, 2017). La double membrane qui constitue l'enveloppe nucléaire était classiquement perçue comme une barrière stricte mais des études montrent que cette frontière peut être contournée, notamment durant les phases de division cellulaire (Remaut et al., 2014).

D'autre part, plusieurs mécanismes peuvent permettre à l'ADN étranger de franchir l'enveloppe nucléaire. Des études montrent qu'une fois dans le cytoplasme,

l'ADN étranger peut s'associer à des protéines possédant un signal de localisation nucléaire (i.e. facteurs de transcriptions ou des histones) et ainsi, être entraîné et franchir les pores nucléaires (Bai et al., 2017). Aussi, les virus sont connus pour être capable d'injecter du matériel génétique dans le noyau cellulaire de leur hôte (Lucic et al., 2021), voire intégrer leur génome dans le cas des rétrovirus (Arnaud et al., 2008; Whitcomb, 1992).

Une fois dans le noyau, l'ADN étranger peut donc être incorporé au génome de l'hôte par l'intervention de virus comme mentionné précédemment ou bien grâce à certaines caractéristiques intrinsèques. L'intégration peut survenir par recombinaison homologue en lien ou pas avec les processus de réparation de l'ADN. Des travaux ont déjà montré que lors de la réparation de l'ADN, des insertions non intentionnelles peuvent survenir (Onozawa et al., 2014). De même, l'activité des transposons (Wicker, 2007) peut aboutir à l'insertion d'ADN étranger, ainsi, les séquences récemment transférées sont souvent bordées de transposons (Parish, 2015). Pour finir, l'insertion peut se faire par recombinaison non-homologue (Milot et al., 1992).

L'ensemble de ces mécanismes illustre comment des gènes peuvent voyager dans les clades de l'arbre du vivant et impacter l'évolution des espèces. Dans la plupart des cas, il est probable que les THG ne subsistent pas dans les génomes hôtes, en raison de la divergence des mécanismes de transcription et de traduction avec l'espèce d'origine. En effet, la plupart des gènes transférés ne sont vraisemblablement pas exprimés par l'hôte et finissent par devenir des pseudogènes. Dans le cas d'une compatibilité permettant l'expression par l'espèce hôte, la probabilité de fixation de ce gène dans la population apparaît faible s'il n'apporte pas d'avantage sélectif, tel que l'adaptation à une nouvelle niche écologique.

4. *THG dans les génomes de nématodes parasites de plantes*

A. Les THG impliqués dans la dégradation de la paroi pecto-cellulosique

1. La paroi pecto-cellulosique et les enzymes de dégradation associées

Les premiers cas de THG décrits dans les génomes de NPP correspondaient à des gènes codant pour des enzymes de dégradation de la paroi pecto-cellulosique entourant et protégeant les cellules végétales. La cellulose est un polysaccharide formé de chaînes de glucose et est produits par les plantes (Sponsler, 1923), les algues (Soni et al., 2021), les Oomycètes (Wu et al., 2019) et certaines bactéries (Ross et al., 1991). Cela en fait le composé organique le plus abondant sur terre. La dégradation de cette macromolécule dépend de différentes enzymes dont des « cellulases » codées par des gènes retrouvés le plus souvent dans les génomes bactériens (Georgelis et al., 2015) ou fongiques (Glass et al., 2013) mais rarement chez les animaux (Shin et al., 2022). Ces enzymes, capables d'hydrolyser des liaisons glycosidiques, sont classées en familles de glycosides hydrolases (GH). Actuellement, l'activité cellulase est retrouvée dans quatorze familles de GH, classées selon leur similarité de séquence en acides aminés.

Bien que la cellulose soit une source alimentaire majeure pour de nombreuses espèces animales, la plupart des omnivores et des herbivores ne produisent pas eux-mêmes de cellulases et ne sont capables de digérer que partiellement la cellulose, et cela uniquement grâce à des micro-organismes cellulosiques vivant dans leur microbiome intestinal. Par exemple, chez les ruminants, les voies digestives comportent des mélanges de bactéries qui dégradent la cellulose dans des conditions anaérobies (Attwood et al., 1996). Pour pénétrer et se déplacer dans les tissus végétaux ainsi que pour se nourrir, les NPP doivent dégrader la paroi cellulaire des plantes. Dès 1963, Victor H. Dropkin a mis en évidence pour la première fois la présence d'enzymes ayant une activité cellulase dans les sécrétions de plusieurs espèces de nématodes à kystes (NK) et de nématodes à galles (NG) (Dropkin, 1963).

2. *Identification de gènes codant pour des cellulases acquis par TH dans le génome d'un organisme animal*

a) Chez les nématodes à kystes

Presque quarante ans plus tard, Smant et collaborateurs ont démontré que la synthèse des enzymes décrites par V. Dropkin chez les NK, se déroulait dans le cytoplasme des glandes oesophagiennes durant le stade larvaire J2 (larves infestantes). Le produit des gènes exprimés dans ces glandes est ensuite sécrété par les nématodes dans les plantes via leur stylet. Les auteurs ont aussi montré que les cellulases synthétisées par les NK appartenaient à la famille GH5 retrouvées généralement chez les bactéries.

Pour cela, les auteurs ont isolé des protéines ayant une activité cellulase à partir de sécrétion des glandes oesophagiennes de deux espèces de NK : *Globodera rostochiensis* et *Heterodera glycines*. Les auteurs ont ensuite séquencé ces protéines et les séquences obtenues ont été utilisées pour concevoir des amorces dégénérées permettant d'amplifier des ADN complémentaires (ADNc) aux ARNm des cellulases. Chez chaque espèce de NK, deux types d'ADNc ont été identifiés correspondant à deux types de protéines de la famille GH5 : un premier comportant un domaine catalytique ainsi qu'un site de liaison à la cellulose, et un second qui possède uniquement un domaine catalytique.

La présence d'introns et de queues poly-A indiquent une signature eucaryote pour les gènes correspondant à ces ADNc. De plus, les analyses d'hybridation *in situ* montrent une expression localisée dans le cytoplasme des cellules oesophagiennes au stade J2 et chez les mâles adultes. Or, jusqu'à présent, aucune étude n'a révélé la présence d'organismes endosymbiotiques eucaryotes au niveau des glandes oesophagiennes subventrales des NPP.

Si ces éléments indiquent que les GH5 appartiennent au génome des NK, les seules séquences homologues identifiées appartiennent à des organismes bactériens dont la bactérie *Erwinia chrysanthemi* (i.e. *Dickeya chrysanthemi)*. Sur la base de ces

observations, les auteurs ont formulé l'hypothèse que ces gènes proviennent d'organismes procaryotes et qu'ils ont été acquis par THG (Smant et al., 1998).

b) …puis, chez les nématodes à galles

Dans un même temps, une étude portant principalement sur le nématode à galles *M. incognita*, Rosso et collaborateurs ont montré que les NG synthétisent aussi des enzymes de type GH5 dans le cytoplasme des glandes oesophagiennes au cours du stade de développement J2 et adulte chez le mâle mais aussi la femelle (Rosso et al., 1999). Dans ce travail, les auteurs ont utilisé les régions conservées des séquences correspondant à des cellulases disponibles dans les librairies de séquences pour concevoir des amorces spécifiques aux gènes codant pour ces enzymes. Les amorces ainsi obtenues ont fourni différents fragments d'ADNc suggérant la présence de multiples copies de gènes qui codent pour des cellulases dans le génome de *M. incognita*. Les ADNc isolés chez cette espèce codent des protéines qui comportent un domaine catalytique et un site de liaison à la cellulose comme l'un des deux ADNc isolés chez les NK.

En revanche, chez les NG, l'expression semble différente au niveau temporel (i.e. stade de développement adulte chez la femelle) mais aussi spatial. En effet, Rosso et collaborateurs ont détecté une activité cellulase au niveau des masses d'œufs. Ce qui suggère également une expression au niveau des glandes postérieures. L'hypothèse est que ces enzymes seraient impliquées dans le processus de ponte chez *M. incognita* (Rosso et al., 1999). Les recherches de similarité menées dans le cadre de cette étude, indiquent une homologie avec les protéines bactériennes dont *Erwinia chrysanthemi* mais ces résultats ne sont pas discutés par les auteurs. Dans la continuité de ces travaux, Béra-Maillet et collaborateurs ont réalisé une caractérisation biochimique plus complète des GH5 de *M. incognita* montrant que les protéines synthétisées par les NPP présentent les mêmes activités biochimiques que les cellulases présentes chez les micro-organismes (Béra-Maillet et al., 2000).

c) Indices d'une acquisition par THG

Chez les NG et les NK, les recherches d'homologie à partir des domaines catalytiques des cellulases indiquent que ces protéines appartiennent à la famille des GH5, une famille habituellement présente chez les bactéries. La classification des cellulases identifiées dans les génomes de NPP dans la même famille que les cellulases bactériennes suggère que les gènes codant pour ces protéines ont évolué à partir d'une même séquence ancestrale, car il est peu probable que ces similarités de séquence soient le fruit d'une évolution convergente (Henrissat et Bairoch, 1993).

La première hypothèse qui pourrait expliquer ces observations est la présence d'un gène codant pour une cellulase dans le génome d'un ancêtre commun aux nématodes et aux bactéries, autrement dit chez le dernier ancêtre commun de l'ensemble du vivant. Cependant, l'absence de ces gènes dans la majorité des génomes des autres nématodes et animaux, des plantes, ou des champignons aurait nécessité des pertes de gènes multiples et massives et constitue une hypothèse peu probable. L'hypothèse alternative est que ces gènes ont été acquis par THG depuis les génomes bactériens dans le génome d'un ancêtre commun aux NPP.

3. Autres THG impliqués dans la dégradation de la paroi pecto-cellulosique

a) Identification et caractérisation individuelle

D'autres fonctions enzymatiques impliquées dans la dégradation de la paroi pecto-cellulosique semblent aussi être codées par des gènes acquis par THG dans les génomes de NPP. Ces enzymes ont été identifiées principalement chez des espèces de NPP endoparasites sédentaires.

La caractérisation fonctionnelle et l'analyse biochimique de ces protéines indiquent qu'elles possèdent des activités xylanases (famille GH30) (Dautova et al., 2001; Mitreva-Dautova et al., 2006), polygalacturonases (famille GH28) (Jaubert et al., 2002), pectate lyases (famille PL3) (Doyle et Lambert, 2002; Popeijus et al., 2000) ou sont des protéines de type expansines n'ayant pas d'activité catalytique mais aidant à

la dégradation ou au remodelage de la paroi cellulaire (Kudla et al., 2007; Qin et al., 2004). D'autres enzymes codées par des gènes acquis par THG semblent présenter une potentielle activité arabinogalactane (famille GH53) et arabinanase (famille GH43) bien que ces dernières n'aient pas encore été caractérisées biochimiquement. Ces enzymes permettent aux NPP de dégrader les chaînes carbonées des glucides composant les tissus végétaux. Les expansines, en revanche, agissent sur les liaisons non covalentes de la paroi cellulaire végétale rendant les composants plus accessibles aux enzymes de dégradation des sucres.

Ces études indépendantes indiquent que, comme chez les champignons et les bactéries pathogènes de plantes, la dégradation de la paroi cellulaire des plantes par les NPP repose sur l'action synergique d'un cocktail d'enzymes et d'autres protéines.

b) L'analyse des génomes complets de NPP révèle l'étendue des THG chez les NPP

Le premier génome d'un animal parasite de plantes publié était celui du nématode à galles *M. incognita* par Abad et collaborateurs en 2008. Bien que ce premier génome fût relativement fragmenté et partiellement incomplet, il permit de retrouver l'ensemble des THG précédemment décrits individuellement mais également de réaliser que nombre de ces gènes avaient subi des duplications après leur acquisition (Abad et al., 2008).

En concordance avec les études précédentes, les auteurs ont retrouvé des cellulases (familles GH5), des xylanases (familles GH30), des polygalacturonases (familles GH28), des pectate lyases (famille PL3) et des expansines. L'annotation manuelle des protéines prédites dans ce génome de *M. incognita* a permis de retrouver les différentes copies de ces gènes pour un total d'une soixantaine de gènes (Abad et al., 2008).

Le travail de Abad et collaborateurs a aussi permis de découvrir, pour la première fois dans le génome d'un animal, des gènes codant pour de possibles arabinanases (famille GH43) et deux possibles invertases (famille GH32). Les

arabinanases sont susceptibles d'être impliquées dans la dégradation de la paroi pecto-cellulosique. En revanche, les invertases ne semblent pas impliqué dans les processus de dégradation de la paroi des plantes mais plutôt au processus de nutrition du nématode. En effet, l'activité invertase consiste en la dégradation du saccharose (sucre circulant dans les plantes) en glucose et fructose, qui sont des sucres utilisables directement par les animaux.

À l'exception de quelques gènes individuellement caractérisés chez certains insectes (Dong Wei et al., 2006), l'acquisition par THG d'un tel arsenal enzymatique n'avait encore jamais été décrite dans le génome d'un animal. Suite à l'identification de ces nombreuses fonctions, des études concernant leur histoire évolutive ont été menées pour mieux comprendre leur origine.

4. *Reconstruction de l'histoire évolutive de ces gènes*

a) Confirmation de l'acquisition horizontale de gènes bactériens ou fongiques

Sur la base de ces observations, Danchin et collaborateurs ont réalisé une analyse phylogénétique systématique des gènes codant des enzymes actives sur les sucres possiblement acquis par THG et codant pour des fonctions biologiques liées au parasitisme des plantes. Dans cette étude, les protéines de nématodes sont systématiquement regroupées avec des protéines d'organismes bactériens ou fongiques. Ces résultats confirment que plusieurs événements de THG indépendants, provenant de différentes sources bactériennes ou fongiques, constituent l'hypothèse la plus probable pour expliquer la présence de ces gènes dans les génomes des nématodes phytoparasites (Danchin, 2010).

Le travail de Danchin et collaborateurs a aussi montré qu'après acquisition via THG, certains des gènes présentent des duplications massives ce qui suggère que les individus possédant plusieurs copies de ces gènes présentaient un avantage sélectif. Les analyses de préférence d'utilisation des codons et de composition en nucléotides G et C de ces gènes acquis par THG, ne montre pas de différence par rapport au

reste des gènes du génome des nématodes à galles, ce qui suggère une acquisition ancienne (Danchin, 2010). De manière concordante, l'analyse de ces séquences a permis de mettre en évidence plusieurs éléments indiquant la domestication de ces gènes par les génomes des NPP tels que la présence d'introns. Danchin et collaborateurs soulignent que l'ensemble de leurs analyses montrent que ces gènes acquis par transferts horizontaux semblent avoir joué un rôle important dans l'acquisition de fonctions biologiques essentielles à ce mode de vie parasitaire.

Étonnamment, les mêmes processus évolutifs sont retrouvés chez d'autres espèces animales parasites de plantes, telle que l'insecte *Phaedon cochleariae*. Ce coléoptère possède des polygalacturonases (PG) lui permettant de digérer la pectine composant les tissus végétaux. Kirsch et collaborateurs ont montré que ces enzymes acquises par THG proviennent de champignons, et qu'elles jouent un rôle essentiel dans les fonctions parasitaires. Les auteurs ont généré, grâce à la technologie Crispr/Cas9, des mutants triple et quadruple-KO pour les gènes PG. L'analyse du phénotype de ces mutants montre une diminution des fonctions biologiques de dégradation de la pectine pour les mutants triple-KO et une perte quasi-totale de cette fonction biologique pour le mutant quadruple-KO. Ces résultats signifient que l'ensemble des fonctions biologiques de dégradation de la pectine chez l'espèce *Phaedon cochleariae* reposent exclusivement sur l'activité biologique de gènes acquis par THG dans le génome de cet organisme et écartent l'hypothèse d'un soutien apporté par des micro-organismes symbiotiques. Ainsi, Kirsch et collaborateurs ont mis en évidence la contribution des gènes acquis par TH à la digestion de la pectine et donc dans les fonctions parasitaires chez cette espèce (Kirsch et al., 2022).

b) Multiples événements de THG chez les NPP

La plupart des THG sont retrouvés chez plusieurs espèces parasites de plantes ce qui suggère une acquisition ancestrale. Cependant, une même fonction peut avoir été acquise indépendamment par THG chez différentes espèces. L'un des

exemples illustrant ce phénomène est le cas des cellulases qui semblent avoir été acquises indépendamment à de multiples reprises dans les clades 2, 10 et 12. Le cas du clade 12 a été décrit en détails précédemment pour *M. incognita*, et nous nous bornerons ici à un bref comparatif avec les clades 10 puis 2.

(1) Chez les NPP de la famille *Aphelenchoididae* (clade 10)

Chez les nématodes à kystes et chez les nématodes à galles (clade 12, Figure 7), la dégradation de la cellulose est due à des cellulases de la famille GH5. Cependant, chez le nématode parasite du pin, *Bursaphelenchus xylophilus* (clade 10b, Figure 7), des glycosides hydrolases présentant une activité cellulase ont été identifiées également mais celles-ci appartiennent à la famille GH45, évolutivement et structurellement distinctes des GH5. Les GH45 sont majoritairement retrouvées dans les génomes de champignons ce qui suggère une acquisition par THG d'origine fongique (Figure 13).

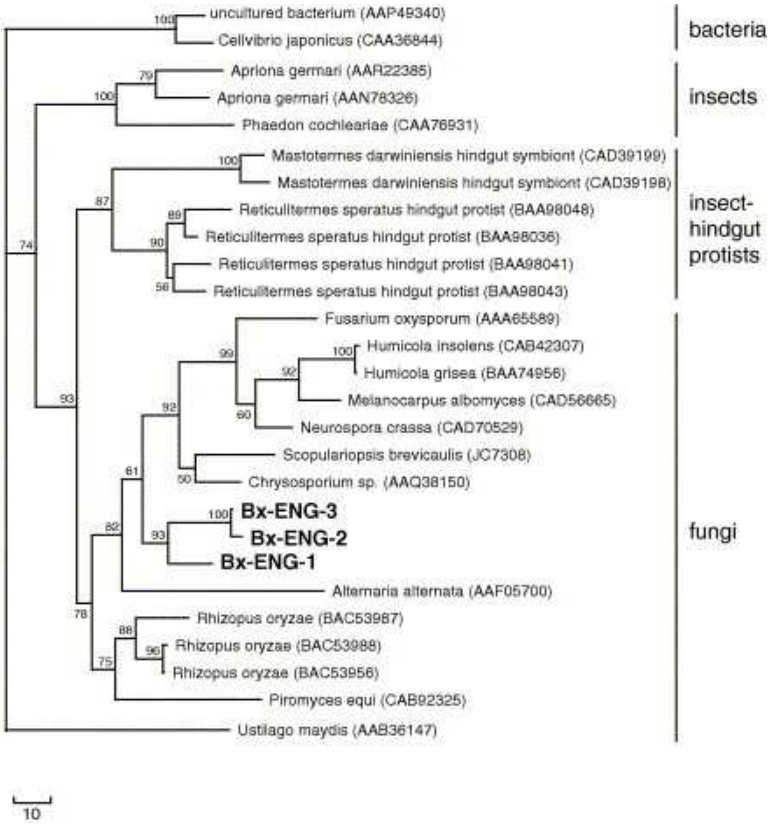


Figure 13. Arbre phylogénétique des cellulases de la famille GH45 présentes dans le génome de Bursaphelenchus xylophilus incluant les protéines homologues fongiques, bactériennes et animales.
Figure modifiée issue de (Kikuchi et al., 2004)

Le nématode *B. xylophilus* est une espèce phytoparasite facultative qui se nourrit aussi de champignons (Kikuchi et al., 2004). L'acquisition de cette fonction pourrait être liée aux processus nutritifs de cette espèce. Une autre étude plus large a mis en évidence spécifiquement la présence de gènes codant des GH45 dans les génomes de différents nématodes de la famille des *Aphelenchoididae* (clade 10b, Figure 7) (Palomares-Rius et al., 2014). Jusqu'alors les cellulases de la famille GH5 et GH45 étaient considérées comme mutuellement exclusives chez les NPP car aucune espèce ne semblait posséder des cellulases des deux familles. Or, récemment, Lai et collaborateurs indiquent la présence à la fois de gènes codant des GH5 et des GH45 dans les génomes de certaines espèces de nématodes appartenant au genre *Aphelenchoides* (Lai et al., 2022).

(2) Chez les NPP de la famille *Longidoridae* (clade 2)

En 2017, Danchin et collaborateurs ont produit des transcriptomes de référence pour deux espèces de NPP, *Xiphinema index* et *Longidorus elongatus*, appartenant au clade 2 selon la classification de Van Megen et collaborateurs (Figure 7). L'analyse de ces transcriptomes a révélé plusieurs dizaines de gènes acquis par THG validés par phylogénie (Figure 14). La caractérisation fonctionnelle et biochimique de ces THG a montré que, chez les NPP du clade 12, les enzymes présentant une activité cellulase provenant d'enzymes appartiennent à la famille GH12 retrouvées généralement chez les archées, Bactéries et Champignons, mais pas chez les Métazoaires (Danchin et al., 2017).

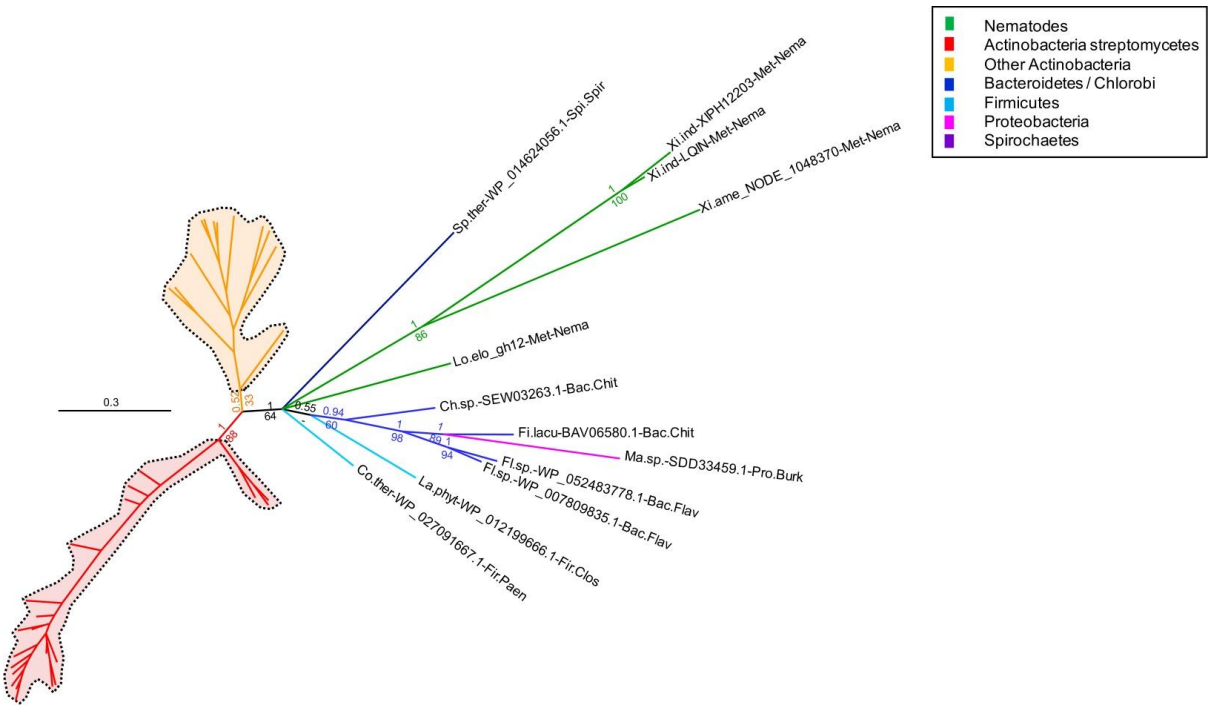


Figure 14. Arbre phylogénétique des cellulases GH12 de Xiphinema index et Longidorus elongatus incluant les protéines homologues bactériennes. Figure modifiée issue de (Danchin et al., 2017)

L'ensemble de ces découvertes indiquent que plusieurs événements de THG ont ponctué l'histoire évolutive des nématodes et que ces événements sont étroitement liés aux multiples émergences du parasitisme (Haegeman et al., 2011).

L'acquisition horizontale de gènes bactériens ou fongiques semble être une alternative évolutive plus directe que l'héritage vertical associé à des mutations pour l'acquisition de fonctions liées à la dégradation de la paroi cellulaire des plantes chez les animaux. Comme vu précédemment, les hypothèses d'un héritage vertical suivi de perte massive ou l'évolution convergente (par néofonctionnalisation ou émergence *de novo*) de séquences similaires à celles trouvées chez les bactéries ou les champignons semblent moins probable qu'un THG.

De nombreux gènes impliqués dans la dégradation de la paroi des plantes ont été identifiés comme provenant de THG. Mais d'autres fonctions biologiques susceptibles d'être impliquées dans le parasitisme semblent aussi avoir été acquises par cette voie.

B. THG impliqués dans les processus de nutrition

Comme on l'a vu précédemment, certains gènes acquis par THG ont des fonctions relatives aux processus nutritifs. Des THG codant pour des invertases de la famille GH32 identifiés pour la première fois chez *M. incognita* (Abad et al., 2008) sont retrouvés également chez *P. penetrans* (Haegeman et al., 2011) et *G. pallida* (Danchin et al., 2016). Les invertases sont connues pour catalyser la conversion du saccharose en glucose et en fructose. Or, le saccharose n'est pas un composant des tissus végétaux mais le sucre-carburant circulant dans les plantes. Cette fonction biologique n'est donc pas liée à la dégradation de la paroi des plantes mais plutôt au processus de nutrition du nématode. Dans une étude portant sur le génome de *G. pallida*, Danchin et collaborateurs ont montré que les GH32 sont exprimées au stade adulte durant le processus d'alimentation au niveau du système digestif de l'animal, ce qui suppose une implication de ces enzymes dans les processus alimentaires des NPP (Danchin et al., 2016). De plus, ces mêmes analyses ont permis de confirmer que ces protéines de la famille GH32 possédaient bien l'activité invertase.

Les autres THG décrits comme impliqués dans des processus nutritifs sont majoritairement impliqués dans la synthèse de plusieurs vitamines B (B1, B5, B6, B7). Graig et collaborateurs ont réalisé un criblage large échelle du génome de référence du nématode *Heterodera glycines.* Cette analyse a révélé la présence de gènes impliqués dans la voie métabolique de la biosynthèse *de novo* de la forme active de la vitamine B6. Cette vitamine interagissant avec l'oxygène est un cofacteur de plus de 140 enzymes différentes (Craig et al., 2008). Les gènes en question sont homologues à des séquences bactériennes, cependant leur analyse a révélé qu'ils contiennent des introns et que leurs transcrits sont polyadénylés. Ces observations indiquent que ces gènes ont été domestiqués par un génome eucaryote.

Certains THG impliqués dans les processus nutritifs sont spécifiques à certains clades. Par exemple, les THG relatifs à la biosynthèse de la vitamine B6 semblent spécifiques au clade des NK alors que d'autres gènes acquis

horizontalement sont retrouvés plus largement chez les *Tylenchida*, tels que les THG impliqués dans la biosynthèse de la vitamine B7.

C. Autres THG et impliqués dans le parasitisme

La première analyse du génome de *M. incognita* avait aussi révélé la présence de chorismates mutases sans que l'origine de ces gènes soit précisée (Abad et al., 2008). Des chorismate mutases avaient précédemment été décrites chez le nématode à galles *M. javanica* (Lambert et al., 1999) et le nématode à kystes *G. pallida* (Jones et al., 2003) et leur plus forte similarité avec des enzymes bactériennes suggérait une acquisition via THG (Haegeman et al., 2011). La chorismate mutase est une enzyme clé dans la biosynthèse des acides aminés aromatiques et des produits apparentés, pouvant agir sur le processus de lignification ou les réponses de défense de la plante dépendant de la tyrosine. Des expériences menées chez le nématode à kyste *H. schachtii* ont montré que l'enzyme du nématode était capable de complémenter l'activité chez une bactérie *E. coli* déficiente pour le gène codant la chorismate mutase (Vanholme et al., 2009). Des THG codant pour cette fonction biologique sont également retrouvés chez *Pratylenchus penetrans* appartenant aussi au clade des *Tylenchina*. Opperman et collaborateurs rapportent aussi la présence de possibles cyanates lyases provenant de THG chez *M. hapla* (Opperman et al., 2008) et leur présence a également été indiquée chez *M. incognita* (Haegeman et al., 2011). Cette enzyme, catalyse le métabolisme du cyanate aboutissant à la formation d'ammoniac, et est généralement retrouvée chez les plantes, les bactéries et les champignons. L'activité biochimique reste cependant à confirmer chez les NPP. La possible cyanate lyase semble rare chez les NPP, bien que des études rapportent la présence de cette enzyme chez les nématodes parasites d'animaux du clade I (Zarlenga et al., 2022) mais aussi chez certains arthropodes parasites de plantes (Wybouw et al., 2014).

Comme nous venons de le voir, les nématodes parasites de plantes ont acquis plusieurs gènes par THG qui ont probablement joué des rôles importants dans l'évolution vers ce mode de vie parasitaire. En outre, de nombreux gènes codant pour des protéines prédites ont été identifiés comme issus de THG dans les génomes de NPP mais leurs fonctions biologiques n'ont pas encore pu être définies (Abad et al., 2008; Danchin et al., 2017; Grynberg et al., 2020; Paganini et al., 2012). Sans information sur la fonction de ces gènes et l'éventuel processus biologique dans lequel ils pourraient être impliqués, il est difficile de déterminer s'ils jouent un rôle dans le parasitisme des plantes. Cependant, la fixation de ces gènes bactériens ou fongiques dans différents clades de nématodes phytoparasites suggère qu'ils ont été sélectionnés au cours de l'évolution possiblement en raison d'un avantage sélectif associé à leur présence (Elling et al., 2009; Wasmuth et al., 2008).

Tableau 1 Synthèse de la présence (Y) / absence (N) des THG décrits dans la littérature dans des génomes de NPP pour les génomes de *Globodera rostochiensis* (*G. ros*) et *Meloidogyne incognita* (*M. inc*). Les fonctions biologiques qui n'ont pas été caractérisées biochimiquement sont indiquées par la mention 'Candidate'.

Bio. Process	Fonction Mol.	Famille de gènes	*G. ros*	*M. inc*	Réf.
CWD	Cellulose degradation	GH5_2 Cellulases	Y	Y	(Béra-Maillet et al., 2000; Danchin et al., 2010; Ledger et al., 2006; Rosso et al., 1999a; Smant et al., 1998)
	Pectin decorations degradation	GH28 Polygalacturonase	N	Y	(Danchin et al., 2010; Jaubert et al., 2002)
		GH43 candidate Arabinanase	N	Y	(Danchin et al., 2010)
	Pectin degradation	PL3 Pecate Lyase	Y	Y	(Danchin et al., 2010; Doyle et Lambert, 2002; Kudla et al., 2007; Popeijus et al., 2000)
	Pectinose / arabinogalactan degradation	GH53 candidate Arabinogalactan endo-1,4-beta-galactosidase	Y	N	(Vanholme et al., 2009)
	Softening of non-covalent bonds	Expansin-like proteins	Y	Y	(Abad et al., 2008; Danchin et al., 2010; Kudla et al., 2007; Qin et al., 2004)
	Xylan degradation	GH30 xylanase	N	Y	(Bakker et al., 2001; Danchin et al., 2010; Mitreva-Dautova et al., 2006)
Detoxification	Unknown	Candidate Cyanate Lyases	Y	Y	(Opperman et al., 2008; Wybouw et al., 2012)
Feeding site induction	Candidate acetyltransferase	NodL - like	Y	Y	(McCarter et al., 2003; Scholl et al., 2003)
Nutrient processing	Degradation of sucrose	GH32 invertase	Y	Y	(Abad et al., 2008; Cotton et al., 2014; Danchin et al., 2016)
	Nitrogen assimilation	Candidate GSI Glutamine Synthase	Y	Y	(Paganini et al., 2012; Scholl et al., 2003)
	Unknown	Candidate PolS Polyglutamate Synthase	Y	N	(Veronico et al., 2001)
	Vitamin B1 biosynthesis	VB1 tenA	Y	N	(Craig et al., 2008)
		VB1 thi4	Y	N	
		VB1 thiD	Y	N	
		VB1 thiE	Y	N	
		VB1 thiM	Y	N	
	Vitamin B5 biosynthesis	VB5 panC	Y	Y	
	Vitamin B6 biosynthesis	VB6 SOR-SNZ	Y	N	
		VB6 SNO	Y	N	
Plant defense manipulation	Conversion of Chorismate into SA	Candidate Isochorismatase	Y	Y	(Bauters et al., 2014)
		Chorismate Mutase	Y	Y	(Jones et al., 2003; Lambert et al., 1999; Vanholme et al., 2009)
Plant defense manipulation	Chitin degradation	GH18 chitinase	Y	N	(Gao et al., 2002)
Unknown	Unknown	Candidate L-threonine aldolase	Y	Y	(Paganini et al., 2012; Scholl et al., 2003)
		Candidate Phosphorybosyl transferase	N	Y	(Paganini et al., 2012; Scholl et al., 2003)

D. L'origine de ces THG dans les génomes de NPP

Parmi tous les possibles mécanismes de transfert introduits dans les chapitres précédents, ceux ayant eu lieu chez les NPP restent inconnus à ce jour, bien que l'on suppose qu'un contact étroit avec des micro-organismes donneurs a été nécessaire, cela pouvant inclure des interactions symbiotiques, pathogènes ou trophiques.

Les types d'interactions nématode - micro-organisme pouvant aboutir à un transfert de gène sont les relations endosymbiotiques ou parasitaires. Quelques travaux mentionnent l'existence de bactéries endosymbiotiques chez les NPP telles que *Wolbachia* ou *Cardinium* qui sont parmi les endosymbiotes ou endoparasites les plus répandus chez les nématodes (Brown et al., 2018; Weyandt et al., 2022). Cependant aucun cas de transfert de gène de ces endosymbiotes vers les génomes des NPP n'a été rapporté jusqu'ici. Notons qu'à part chez le NPP du Clade 2 *Xiphinema americanum*, l'existence même d'ensosymbiontes chez les autres NPP reste à démontrer (Palomares-Rius et al., 2014).

En revanche, l'analyse des protéines homologues aux THG identifiés chez les NPP suggère que ces gènes proviendraient de micro-organismes telluriques parasites ou symbiotes de plantes. Certains micro-organismes vivant dans le sol ou la rhizosphère interagissent avec les plantes (Kundu et al., 2022) et possèdent un répertoire de gènes très varié dont certains sont susceptibles d'être utiles à un mode de vie parasitaire (Fierer, 2017; Parks et al., 2017; Torsvik et Øvreås, 2002). Les organismes pathogènes aussi bien que symbiotiques ont développé des fonctions biologiques impliquées dans le contournement des défenses de la plante ou le métabolisme des produits végétaux. Ces organismes présentent un large panel de gènes codant pour des protéines impliquées dans des fonctions biologiques qui ont pu contribuer à l'adaptation des nématodes à un mode de vie parasitaire.

Les résultats des différentes études réalisées jusqu'à aujourd'hui soutiennent cette hypothèse. Par exemple, les gènes codant pour des glycoside hydrolases de la famille 32 présentant une activité invertase chez les NPP sont homologues à ceux

présents chez les bactéries du genre Rhizobacteria, qui sont des bactéries symbiotiques de légumineuses et fixatrices d'azote (Danchin et al., 2016, Haegeman et al., 2011). Cependant, l'ensemble des analyses portant sur les THG chez les NPP réalisés jusqu'à aujourd'hui utilisait des librairies de séquences de références généralistes constituées principalement de données provenant de micro-organismes cultivés en laboratoire. Cela a donc probablement limité la possibilité d'identifier des homologues des THG chez les micro-organismes du sol car beaucoup d'entre eux ne sont soit pas cultivables en laboratoire soit complètement inconnus à ce jour. L'essor de la métagénomique a depuis démontré que ce type de librairies de séquences généralistes n'est pas du tout représentatif de la diversité génétique des communautés microbiennes environnementales (Parks et al., 2017).

5. *Objectifs*

Comme nous l'avons vu, les espèces parasites de plantes constituent un groupe polyphylétique au sein des nématodes. Ce mode de vie parasitaire a donc vraisemblablement évolué indépendamment à de multiples reprises. Par ailleurs, les analyses génomiques ont révélé la présence de gènes fongiques et bactériens acquis par transferts horizontaux chez l'ensemble des espèces parasites de plantes étudiées. La caractérisation fonctionnelle et biochimique des protéines codées par ces gènes indiquent que certains d'entre eux sont impliqués dans des fonctions essentielles à ce mode de vie parasitaire. Ainsi, les THG ont probablement joué un rôle important dans les multiples émergences du parasitisme des plantes chez les nématodes.

Il est supposé que ces gènes proviennent de micro-organismes telluriques qui étaient présents dans l'environnement naturel de ces nématodes. Jusqu'à présent, les analyses de THG ont été réalisées avec des librairies de séquences généralistes. Or, les études métagénomiques ont montré que ces ressources ne sont pas représentatives de la diversité génétique des environnements naturels. Les ressources produites avec ces approches métagénomiques offrent l'opportunité de constituer une librairie de séquences de références plus représentative du pool de gènes présent dans l'environnement dans lequel évoluent ces parasites.

D'autre part, la majorité des études portant sur les gènes acquis par TH chez les nématodes parasites de plantes sont centrées sur une espèce ou un genre de nématodes parasites de plantes; ou bien portent sur un ensemble de nématodes mais se restreignent à certaines familles de gènes. Ainsi, plusieurs questions restent en suspens concernant l'origine et l'histoire évolutive de ces gènes.

Au cours de cette thèse, l'objectif était d'évaluer l'étendue des gènes acquis par THG chez les NPP et de préciser l'origine et l'histoire évolutive de ces gènes. Pour cela, j'ai réalisé une étude à large échelle de THG sur les génomes de différentes espèces du clade *Tylenchina* en utilisant une librairie de référence intégrant des données métagénomiques de sol.

Le premier objectif était de constituer une librairie protéique représentative de la diversité des micro-organismes telluriques en intégrant les protéines provenant de plus de 6 800 métagénomes de sol et leurs métadonnées associées. Les micro-organismes eucaryotes étant souvent mal représentés dans les données métagénomiques, l'un des défis était d'améliorer la prédiction des protéines provenant de ce type de micro-organismes. Pour cela, j'ai développé une méthode bioinformatique permettant de détecter les contigs eucaryotes puis de prédire les gènes à partir de données métagénomiques de sol. J'ai ensuite utilisé une méthode de dernier ancêtre commun basée sur des résultats de recherche d'homologie afin d'assigner une annotation taxonomique à ces gènes. Enfin j'ai combiné ces données avec celles d'autres ressources métagénomiques et la banque de données nr du NCBI afin de générer une librairie de séquence la plus riche possible et améliorant la représentation de la diversité de gènes présents dans les sols.

Mon second objectif était d'utiliser cette ressource pour évaluer de manière la plus exhaustive possible, la contribution des événements de transferts horizontaux à la composition du génome des NPP. Afin d'obtenir une vision intégrative du phénomène de THG chez les NPP, j'ai décidé d'inclure plus d'une quinzaine d'espèces de NPP couvrant une diversité de modes de parasitisme, dont sédentaires, migrateurs, endoparasites ou semi-endoparasites. L'inclusion d'une telle diversité de NPP devrait également permettre de déterminer quels événements de THG semblent ancestraux à différentes espèces par rapport aux événements spécifiques à un genre ou une espèce. J'ai ainsi comparé l'ensemble des protéines prédites chez ces NPP à la librairie de séquences enrichie en données métagénomiques des sols puis utilisé des

méthodes phylogénétiques afin de détecter de manière la plus robuste possible les événement de THG.

Une fois un inventaire exhaustif et robuste des THG chez les PPN déterminé, mon troisième objectif a été d'étudier leurs possibles fonctions afin d'évaluer l'impact des THG sur la biologie des NPP. Cela m'a permis d'une part de confronter les résultats de cette étude à la littérature mais aussi d'évaluer le potentiel de découverte de nouvelles fonctions qui pourraient avoir été acquises par cette voie, je me suis ainsi intéressée aux domaines fonctionnels présents dans les protéines codées par ces gènes.

Le dernier objectif concernant cette partie du projet était de déterminer si les données de métagénomiques de sols nous permettaient d'en apprendre plus quant à l'identité des donneurs et si la fouille de ces données avait permis de révéler des cas qui auraient été impossibles à détecter en l'absence de données métagénomiques des sols.

Enfin, un objectif plus appliqué durant la phase d'exploitation des données métagénomiques a été d'extraire des informations concernant la distribution des champignons mycorhyziens à arbuscules afin d'alimenter les bases de données de la société mycophyto.

Chapitre I

Chapitre 1: Amélioration de la représentation des micro-organismes eucaryotes dans les métagénomes de sols

1. Contexte

Depuis le développement des approches métagénomiques, le nombre d'analyses de ce type ne cesse de se multiplier, et les données générées dans le cadre de ces études sont généralement mises à disposition sur des serveurs publics. L'une des plus grandes ressources métagénomiques publiques est la base de données IMG/M du *Joint Genome Institute* (JGI). Cette ressource comporte des échantillons provenant de tous types de microbiomes et issus de prélèvements à travers le monde entier, bien qu'une grande proportion de ces données proviennent des USA (Figure 15). Pour obtenir une représentation plus exhaustive de la biodiversité du sol, nous avons intégré l'ensemble des jeux de données métagénomiques provenant du séquençage d'échantillons de terre ou de rhizosphère qui étaient disponibles publiquement au début de cette étude sur les serveurs du JGI.

Figure 15 Carte des zones d'échantillonnage des données métagénomiques terrestres disponibles sur le site IMG/M du JGI en janvier 2021.

Les serveurs publics du JGI ou d'autres laboratoires proposent le stockage des données brutes mais ils mettent aussi à disposition les résultats d'analyses issues des pipelines dédiés au traitement de ces données brutes. Ces pipelines intègrent généralement les étapes d'assemblage des lectures, de prédiction des gènes, ainsi que les annotations fonctionnelles et taxonomiques. Ces grandes étapes sont communes aux pipelines, cependant les outils utilisés d'un serveur à l'autre mais aussi selon les versions d'un même pipeline peuvent être assez différents (Figure 16 A). De même, les protocoles d'extraction de l'ADN peuvent varier d'un laboratoire à l'autre. Et enfin, même si la majorité des données proviennent de séquençage à lectures courtes, la technologie de séquençage choisie peut impacter la qualité des données (Figure 16 B). Afin de constituer une librairie de protéines du sol de référence, j'ai dû au préalable réaliser plusieurs analyses pour contrôler la qualité et la fiabilité des protéines prédites, mais aussi leurs affectations taxonomiques fournies par le pipeline du JGI.

L'un des points critiques concernant la qualité globale des données pré-analysées disponible sur le JGI était la représentation des micro-organismes eucaryotes. En effet, la majorité des études métagénomiques se concentre sur les

procaryotes et en particulier les bactéries qui constituent la plus importante part des communautés microbiennes si on considère le nombre d'individus. Cependant, en termes de biomasse, les micro-organismes eucaryotes constituent une part presque aussi importante que les organismes procaryotes (Fierer, 2017). La composition et la diversité des micro-organismes eucaryotes dans le sol devraient être plus élevées et différentes de celles des autres écosystèmes, mais sont encore mal connues (Grossmann 2016, Mahé 2017). De plus, ces micro-organismes présentent une large diversité et jouent des rôles clés dans les écosystèmes.

Des études récentes estiment que leur diversité sur terre est « quasi impondérable » en dénombrant des centaines de milliers de groupes taxonomiques dans la zone euphotique de l'océan. Affichant différents modes de vie, ces micro-organismes eucaryotes participent pleinement aux chaînes trophiques et contribuent aux cycles du carbone (Jardillier et al., 2010) et de l'azote (Woehle et al., 2018). Ils interagissent autant avec les autres membres des communautés microbiennes (Scheu et al., 2005) qu'avec les organismes multicellulaires végétaux (Bonfante et Genre, 2008) et animaux (Mohammad Rahimi et al., 2022). Malgré cela, ces micro-organismes sont négligés par les procédures d'analyse métagénomique automatisées, aboutissant à leur sous-représentation (West et al., 2017).

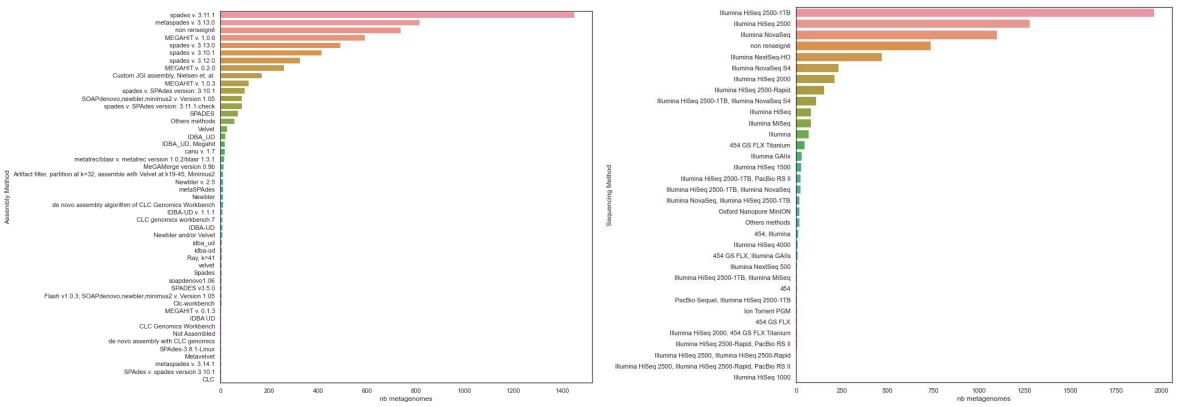


Figure 16 Métadonnées, associées aux métagénomes terrestres et rhizosphère provenant du JGI, indiquant les (A) Méthodes d'assemblages et les (B) Méthodes de séquençage appliquées aux échantillons.

La sous-représentation des organismes eucaryotes dans les données métagénomiques est due à une combinaison de plusieurs facteurs tels que de faibles densités de populations dans les échantillons de sol, induisant une moindre représentation dans les librairies de séquences de référence à laquelle il faut ajouter l'utilisation d'outils inadaptés pour le traitement de ce type de données. Par exemple, la prédiction des gènes réalisée par les procédures automatiques telles que proposées par le JGI utilise uniquement l'outil bio-informatique Prodigal (Hyatt et al., 2010), qui est un outil qui a été spécifiquement dédié à la prédiction de gènes procaryotes. Or, les structures génomiques sont très différentes entre ces organismes procaryotes et eucaryotes plus complexes. Même si la longueur moyenne des protéines est assez proche (e.i. bacteria : 270 aa, archaea : 242 aa et eucaryotique : 353) (ref. nevers), les gènes eucaryotes sont en général beaucoup plus longs en raison de la présence d'introns et des processus d'épissage (Figure 17). Ainsi, l'identification de gènes codant pour les protéines d'organismes eucaryotes est plus difficile, et les outils tels que Prodigal ne sont pas adaptés.

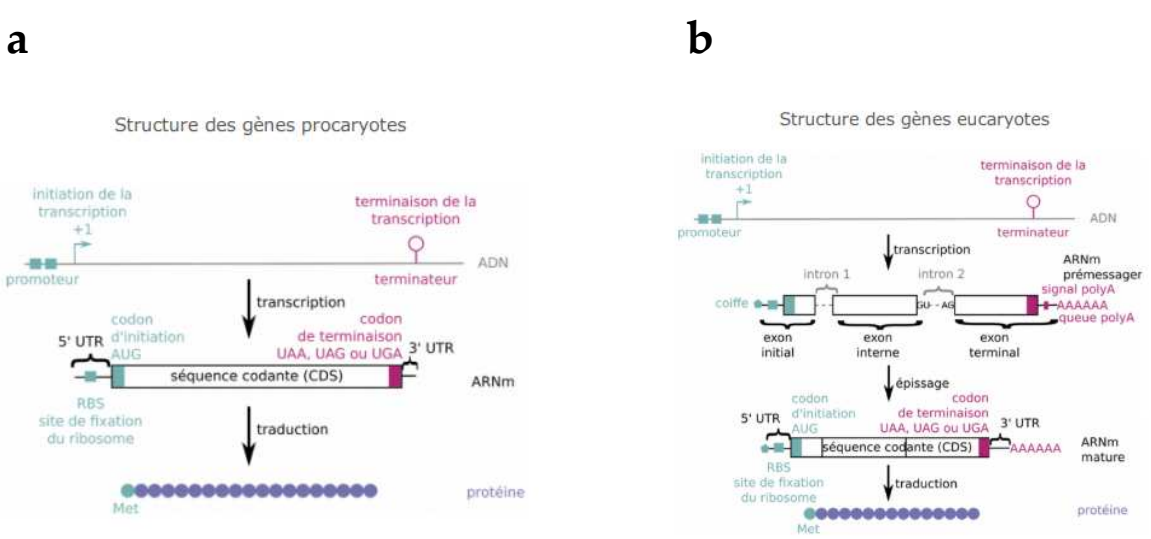


Figure 17 Schémas de la structure des gènes (A) procaryotes et (B) eucaryotes.

Cependant, ces micro-organismes sont particulièrement importants dans le cadre de nos recherches car, comme nous l'avons vu précédemment, certains gènes acquis par TH chez les NPP proviennent de micro-organismes eucaryotes, comme par exemple les gènes codant pour des cellulases de la famille GH45 chez *B. xylophilus*. Nous pouvons préciser qu'il s'agit aussi des micro-organismes du sol par lesquels Mycophyto est particulièrement intéressée car actuellement les bio-stimulants proposés par cette société sont composés de Champignons Mycorhiziens à Arbuscules (CMA). Il était donc crucial pour nous d'obtenir une représentation optimale des micro-organismes eucaryotes telluriques dans notre base de données.

Pour remédier à cet écueil, j'ai du (i) identifier les séquences eucaryotes, (ii) prédire les gènes et protéines sur ces séquences en utilisant des outils adaptés aux eucaryotes et (iii) sur la base des prédictions de protéines améliorées, assigner une annotation taxonomique plus complète et plus fiable. De cette manière, nous avons reprédit plus de 8 millions de protéines eucaryotes de manière plus complète au sein de plus de 6,800 métagénomes de sols et celà avec une annotation taxonomique améliorée. De plus, j'ai prédit 300,000 protéines présentes sur des fragments de génomes eucaryotes qui n'ont pas d'homologie ailleurs dans les bases de données de séquences et semblent donc complètement nouvelles. L'ensemble de ce travail est présenté plus en détail dans l'article suivant.

2. *Article*

scientific data

OPEN

DATA DESCRIPTOR

Improvement of eukaryotic protein predictions from soil metagenomes

Carole Belliardo[1,2] ✉, Georgios D. Koutsovoulos[1], Corinne Rancurel[1], Mathilde Clément[2], Justine Lipuma[2], Marc Bailly-Bechet[1,3] & Etienne G. J. Danchin[1,3] ✉

During the last decades, metagenomics has highlighted the diversity of microorganisms from environmental or host-associated samples. Most metagenomics public repositories use annotation pipelines tailored for prokaryotes regardless of the taxonomic origin of contigs. Consequently, eukaryotic contigs with intrinsically different gene features, are not optimally annotated. Using a bioinformatics pipeline, we have filtered 7.9 billion contigs from 6,872 soil metagenomes in the JGI's IMG/M database to identify eukaryotic contigs. We have re-annotated genes using eukaryote-tailored methods, yielding 8 million eukaryotic proteins and over 300,000 orphan proteins lacking homology in public databases. Comparing the gene predictions we made with initial JGI ones on the same contigs, we confirmed our pipeline improves eukaryotic proteins completeness and contiguity in soil metagenomes. The improved quality of eukaryotic proteins combined with a more comprehensive assignment method yielded more reliable taxonomic annotation. This dataset of eukaryotic soil proteins with improved completeness, quality and taxonomic annotation reliability is of interest for any scientist aiming at studying the composition, biological functions and gene flux in soil communities involving eukaryotes.

Background & Summary

Soil-dwelling microorganisms play essential biological functions related to human and Earth health in both managed and natural ecosystems[1]. In recent years, the rise of metagenomics has expanded our understanding of the genetic diversity of microorganisms in many different complex environments, including soil and plant-associated microbiomes[2]. Metabarcoding and shotgun metagenomic sequencing have highlighted the high diversity of microbial communities and allowed the discovery of previously unknown microorganisms[3,4]. Recent efforts have focused on the *de novo* assembly of bulk metagenomic sequencing reads into metagenome-assembled genomes (MAGs) or contigs, uncovering the genetic content and informing on the molecular functions of these microorganisms[5–7].

The soil is arguably one of the most complex microbiome due to the extremely high diversity of organisms, their complex inter-kingdom interactions and the wide spectrum of environmental conditions observed between samples. In comparison, the human gut microbiome is more homogeneous among individuals due to more stable physiological conditions. Therefore, the soil contains many microbial guilds which cover all different superkingdoms of life with disparate metabolic abilities[8]. Most metagenomic studies are focused on bacteria, which dominate microbiome in number of individuals, although eukaryotes often account for a comparable biomass in soils[2]. The composition and diversity of eukaryotic microorganisms in soils are expected to be higher than, and different, from other ecosystems but are still mostly unknown[9–11]. Moreover, eukaryotic soil microorganisms fulfill essential functions in ecosystems, mainly by participating in the biochemical balance[12] and nutrient cycling[13]. They also affect the biodiversity and health of macro-organisms constituting fauna and the flora. Some eukaryotes are pathogens of plants or animals, and can cause tremendous health or economic damages[14]. In contrast, some others are beneficial such as mycorrhizal fungi which live symbiotically with 90% of the vascular plants on Earth[15]. The mutualistic interactions of plants with eukaryotic microorganisms from the rhizosphere provide them nutritive and protective benefits, giving those fungi a strong agronomic and environmental interest[16–19].

Despite their prime importance in diverse processes, soil eukaryotes are neglected and not well represented in public metagenomic data. Previous studies have highlighted the poor representation of eukaryotes

[1]Institut Sophia Agrobiotech, Université Côte d'Azur, INRAE, CNRS, Sophia Antipolis, France. [2]MYCOPHYTO, 540 Avenue de la Plaine, 06250, Mougins, France. [3]These authors contributed equally: Marc Bailly-Bechet, Etienne G. J. Danchin. ✉e-mail: Carole.Belliardo@inrae.fr; Etienne.Danchin@inrae.fr

in standard metagenomics analyses in different environmental samples and proposed strategies to mitigate this under-representation[20,21]. The largest publicly available resource for soil metagenomes is the Integrated Microbial Genomes & Microbes (IMG/M) database of the Joint Genome Institute (JGI)[22]. In this resource, standard pipelines are used to assemble and annotate contigs and genomes from environmental metagenomic shotgun reads. One major limitation concerns the eukaryotic component of these soil metagenomes. Indeed, the gene prediction tool used by default for all contigs assembled from metagenomes is Prodigal[23], a software tailored for prokaryotes. However, gene structures and features are different in eukaryotes, and using prokaryotic tools to predict eukaryotic genes can lead to incomplete, erroneous and discontinuous gene sequences, and hence proteins: a trivial example is that no intron can be predicted by Prodigal. These procedures make sense given the volume of metagenomic data processed by IMG/M, but, as a consequence, eukaryotic proteins are neglected in these soil microbiome data, with a risk of being truncated and assigned an unreliable taxonomic annotation. These suboptimal sequences and taxonomic annotations then negatively impact any research on the eukaryotic component of the soil.

To circumvent this problem, we have constituted a dataset of 6,872 soil microbiomes comprising 7.9 billion contigs and identified eukaryotic contigs using a k-mer based approach. On the identified eukaryotic contigs, we re-predicted ca. 93 million genes and proteins using annotation methods tailored for eukaryotes. We re-assigned taxonomic information to these proteins based on a last common ancestor (LCA) approach from homology search against the NCBI's nr library. This allowed identifying 8 million eukaryotic proteins and more than 300,000 orphan proteins located on eukaryotic contigs and lacking homology in public protein libraries, representing a potential for new discoveries. We show that the newly predicted proteins are longer and constitute a more comprehensive representation of the pool of eukaryotic proteins in the soil.

This new dataset improves eukaryotic protein sequence quality and completeness, as well as the reliability of the taxonomic information, and represents a unique resource to decipher and study the pool of eukaryotic proteins present in the soil.

Methods

Data collection. We used publicly available assembled metagenomic data from shotgun sequencing reads of the IMG/M database of the JGI[22]. We collected metagenomes of 5,988 'Terrestrial' samples in the environmental metagenomes category and 884 plant-associated metagenomes in the host-associated category, Fig. 1 (available data 2020, October; Supplementary Data[24]). Most of the datasets were unrestricted from use, according to the JGI policy; the authors of a few datasets (see Acknowledgements) that were still under use-restriction kindly authorized us to re-use their data, including two published in the literature[25,26]. The data acquisition was performed via the IMG/MER Cart genome portal. For each metagenome, the JGI provides a set of files from pre-computed analyses that are useful to sort, filter and describe data. Because we anticipated substantial differences in the relative proportions of eukaryotic species present in the terrestrial and the host-associated categories, these two datasets were processed separately to minimize potential biases. For a more convenient processing of the massive amount of data, the metagenomes from terrestrial samples were splitted in two batches; 'Terrestrial 1' contained 3,601 environmental metagenomes added between December 2009 and January 2019 and 'Terrestrial 2' contained 2,387 metagenomes added between February 2019 and August 2020.

Data curation and quality control. Starting from assembled contigs, we combined all genomic fasta files by datasets and obtained 6 and 1.9 billion contigs from terrestrial and plant-associated categories, respectively. The length distribution of assembled contigs is highly heterogeneous between metagenomes due to variation in sequencing technologies, experimental protocols, pipeline version used and biological features. Probably because most data initially consisted of short sequencing reads, half of the contigs were shorter than 296 bp (Table 1). These short contigs increase the volume to be processed and are unlikely to contain complete genes, hence providing no more information on gene diversity[27]. Thus, we filtered data on assembly length, and kept contigs at least 1 kb long or containing at least three genes predicted by Prodigal in JGI files. Only 763 million contigs (10%) passed this filter and were retained for further analysis (Fig. 1). These remaining contigs were distributed in 6,610 metagenomes (Fig. 1): hence data from 262 starting metagenomes were entirely removed due to a too high level of fragmentation. This quality filtering drastically reduced the dataset volume and ensured we only worked with contigs on which complete eukaryotic genes have a chance to be predicted.

Detection of contigs from eukaryotic organisms. The JGI provides taxonomic information for genes predicted using Prodigal, which is not suitable for eukaryotic genes. Moreover this taxonomic information has been transfered solely from that of the best BLAST hit, which can be misleading. Thus, this information cannot be used to identify potential eukaryotic contigs and no further taxonomic information is provided for contigs. Therefore, we scanned all the contigs and identified those from eukaryotic origins using Kraken2[28], a taxonomic classification tool based on exact kmer matches, designed to process in a fast and sensitive way large data sets such as those from metagenomics analyses. Among the taxonomic classifiers dedicated to metagenomic data, we selected Kraken2 because it provides taxonomically labeled contigs and it is designed to work on reads but can also process contigs. As a consequence, this software maintains a good sensitivity on short sequences, representing an ideal choice in our case. Indeed, as indicated in Table 1, our data mainly contains short contigs with an average size of 1.9Kb, which would be sub-optimal for usage with a contig-centered software such as Eukrep[20],that performs better on contigs at least 3Kb long, as mentioned by the authors. As a reference database, we combined all RefSeq libraries of complete genomes [Archaea, Bacteria, Plasmid, Viral, Human, Fungi, Plant, Protozoa][29], complemented by the NCBI's nt library and ran Kraken2 with default parameters. This allowed assigning taxonomic information to 82% of contigs, among which 113 million were classified with a eukaryotic taxonomic identifier 'TaxID' (Fig. 1).

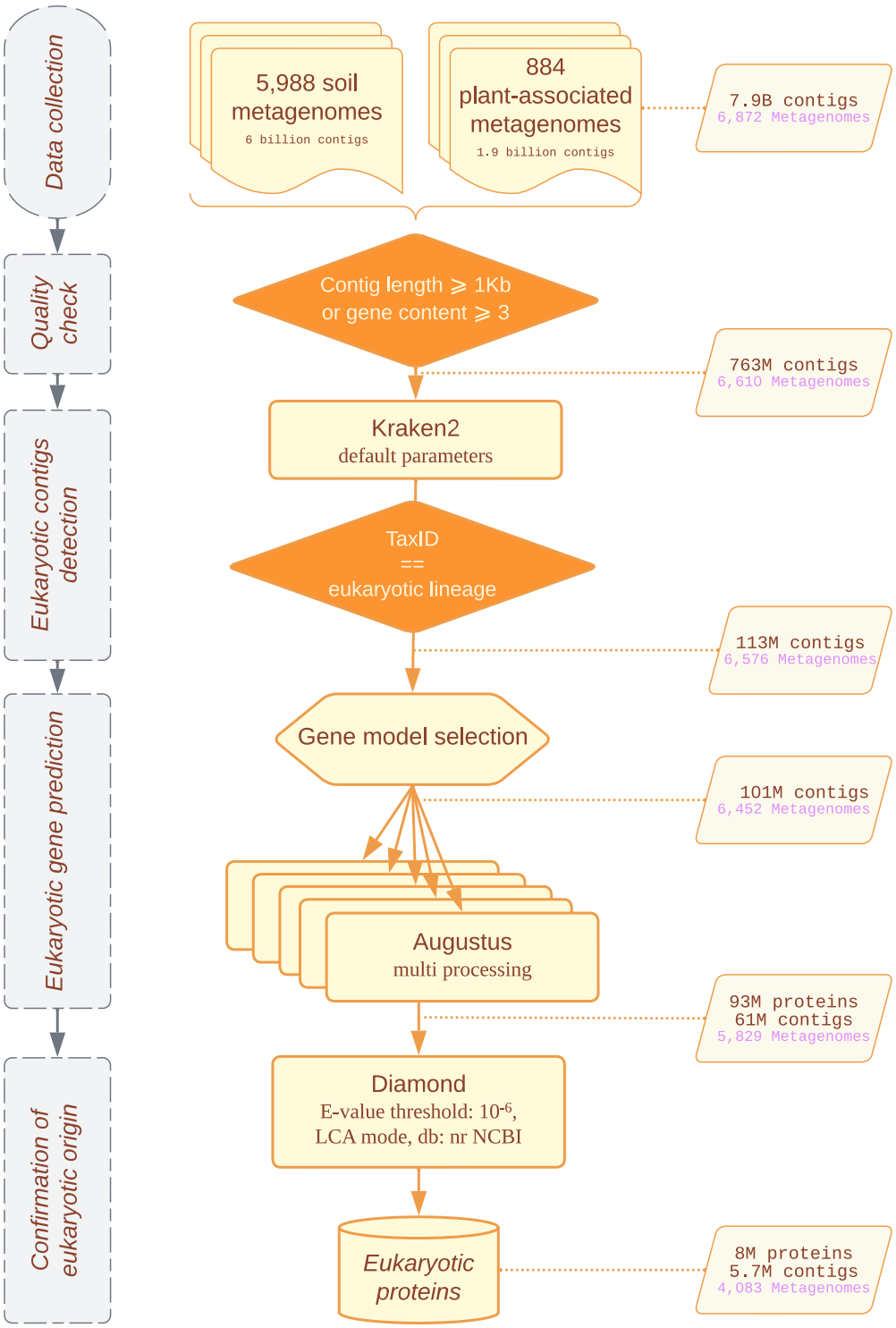


Fig. 1 Our eukaryotic protein prediction pipeline from soil metagenomic contigs to a final dataset of taxonomically annotated proteins with contigs, proteins and metagenomes number at each step.

Eukaryotic gene prediction. For all contigs identified as eukaryotic by Kraken2, we used Augustus (v3.3), a software dedicated to *de novo* eukaryotic gene prediction[30]. The gene structure is complex in eukaryotes and changes across species[27]. Thus, Augustus provides *ab initio* models for 73 different species (Fig. 2) and one must be selected to perform gene prediction. Due to the conservation of genomic features across closely related organisms, we assigned, to each eukaryotic contig, a model based on its Kraken2 taxonomic annotation. Note that this model selection step does not aim at a definitive taxonomic annotation; here we used a sensitive approach to predict as accurately as possible putative eukaryotic genes that will then be filtered by a more selective

Data	Metric	Min	Mean	Median	Max
Raw	Number of contigs per metagenomes	1	1,160,141	294,105	39,582,895
	Contig length (pb)	3	497	296	5,373,015
	Number of genes per contig	0	1	1	5,459
Filtered	Number of contigs per metagenomes	1	115,615	22,307	3,625,639
	Contig length (pb)	1,000	1,985	1,350	5,373,015
	Number of genes per contig	1	3	2	5,459

Table 1. Metrics to assess the contiguity of the 6,872 'Terrestrial' and 'Plant-associated' metagenome-assembled genomes datasets from the IMG/M server of the JGI including the number of proteins predicted by Prodigal from IMG/M.

homology-based taxonomic annotation approach at the protein level. Selection of the phylogenetically closest model for gene prediction on each contig was done using a custom python script[31] which functions as follows:

- First, we browsed the 73 model species tree from the leaves to the root assigning a non-ambiguous parental taxonomic term to each model species as long as no bifurcation with a branch containing another model species was found (Fig. 2). For example, in plants, *Arabidopsis thaliana* is the sole representative of the Brassicales; so the Brassicales parental term was associated with the *A. thaliana* model. Consequently, we used the *A. thaliana* Augustus model for all eukaryotic contigs assigned with a taxonomic ID belonging to the Brassicales branch. Similarly, *Homo sapiens* is the only representative of mammals, so any contig identified by Kraken2 as a mammalian organism will be assigned the *H. sapiens* model.

At this point, an Augustus gene prediction model could be assigned to 7.1% (ca. 8 million) of contigs. The rest of the contigs (ca. 105 million) could not be assigned an unambiguous closest model species because they belonged to a bifurcating branch in the tree leading to several equally close model species.

- Therefore, in a second step, for all these remaining eukaryotic contigs, we selected among the children branches the most frequently assigned model in the whole dataset the contig belongs to (i.e. Plant-associated, Terrestrial 1 or Terrestrial 2) at the previous step (first pass). To continue the previous example, the next more ancestral branch in the phylogeny of Brassicales is the clade 'Malvids' that displays a polytomy of eight children branches of which only two contain an Augustus model species (Malvales and Brassicales). Hence, no model could be unambiguously assigned to contigs with a Malvids taxonomic ID other than Malvales or Brassicales. Therefore, all contigs from other Malvids orders are processed with the most frequently assigned species model for each of the three datasets (Fig. 3). For example, they are processed with the cocoa gene model (Malvales, *Theobroma cacao*) in the dataset Terrestrial 1, or the *Arabidopsis thaliana* gene model in the Plant-associated and Terrestrial 2 datasets (Fig. 3). The distribution of contigs across the models is available in Supplementary Data, Fig. 1[32].

Overall, our pipeline allowed assigning an Augustus model to ca. 101 million possibly eukaryotic contigs (Fig. 1). The most assigned ones were Metazoa and Viridiplantae models, with respectively 49% and 44% of contigs in plant-associated metagenomes and 76% and 16% in terrestrial data. In both datasets, we assigned fungal models to 6.5% of contigs (Supplementary Data, Table 1[33]); and the majority of other contigs were assigned to SAR, Discoba or Rhodophyta models. Although these last taxonomic groups were assigned at a relatively low proportion, this still corresponds to tens or hundreds of thousands of contigs. Unsurprisingly, the less assigned are gene models of aquatic animals such as some benthic animals, sharks, or also lamprey models. At this point, we could not assess whether the numerous assignments to metazoan and plant models came from mis-annotated contigs or contamination, therefore further analyses were performed after gene and protein prediction.

Then, once a model species has been assigned to contigs we ran the eukaryotic gene predictor Augustus[30], with default parameters, which allowed predicting 93 million protein-coding genes (Fig. 1). The number of proteins predicted per contig ranges from 1 to 410 with 2 protein predicted per contig on average for all datasets together. Consistent with the model assignment across kingdoms, the highest numbers of proteins were predicted for contigs assigned to Metazoan and Viridiplantae Augustus models. Moreover, we predicted 8.7 million proteins with Augustus fungal models and 1.8 million with different protist models (Fig. 4a).

Confirmation of eukaryotic origins and improvement of the taxonomic information. To filter-out false-positive eukaryotic classification and assign a more reliable taxonomic annotation to the proteins predicted by our pipeline than simple inheritance from the Kraken2-based contig annotation, we used the last common ancestor algorithm of Diamond[34]. The homology search was run at the protein level with an E-value threshold of 10^{-6} and using the January 2020 release of the NCBI nr database[29] as protein reference. In the LCA mode, Diamond will assign an NCBI taxonomic identifier (i.e. TaxID) based on the last common ancestor of all the hits with a score not diverging by more than 10% from the best hit score. Using an LCA approach constitutes a substantial gain in taxonomic annotation reliability compared to approaches based on the best BLAST hit alone, this single best hit being potentially mis-annotated itself, or sharing only low identity with the query sequence. This LCA approach is usually employed for taxonomic assignment of sequences distantly related to those of known organisms present in public sequence libraries, such as ancient or actual metagenomic data[28,35–38].

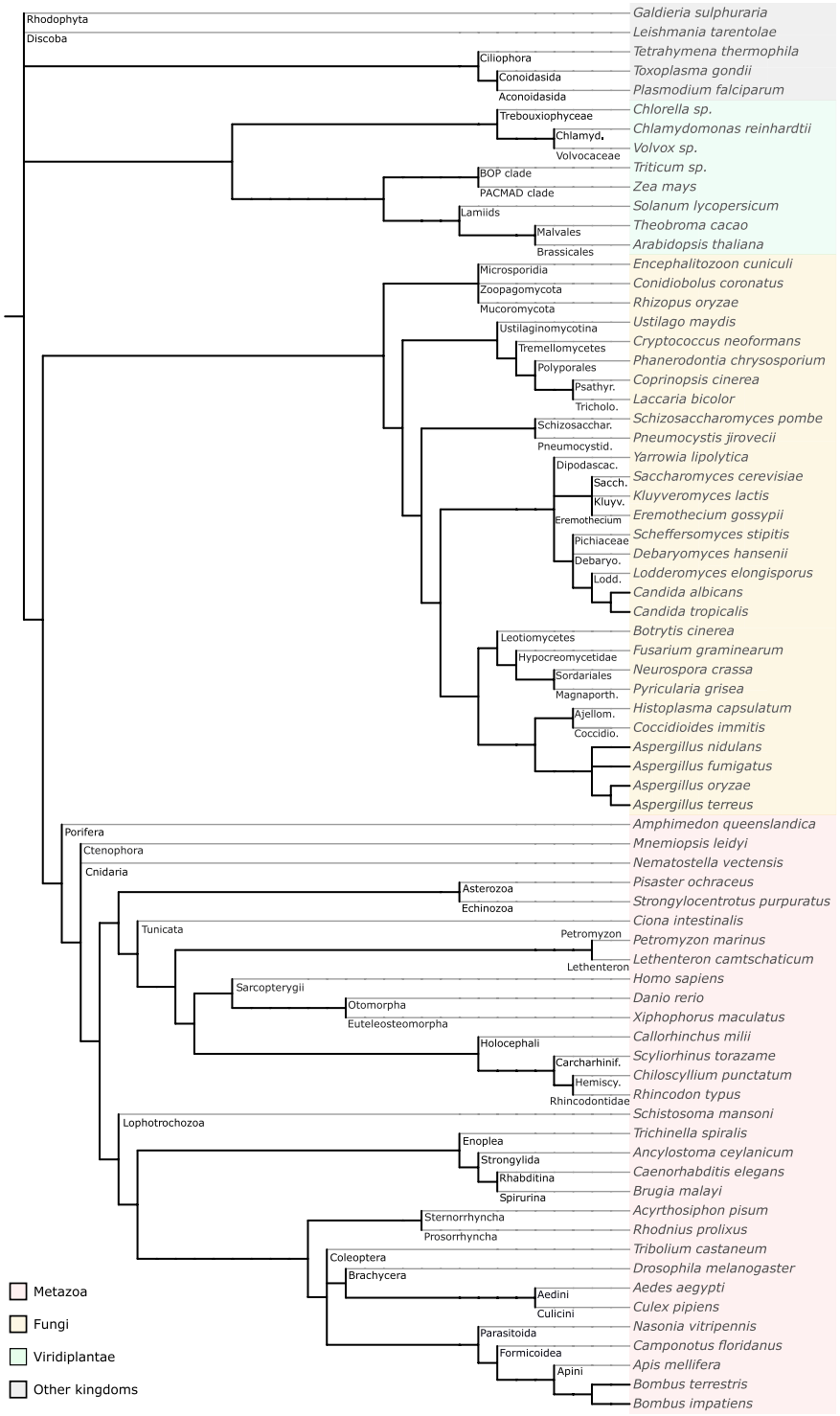


Fig. 2 Phylogenetic tree of Augustus *ab initio* models showing the deeper taxonomic nodes used in the first step of the contig model selection.

Consequently, the improved quality and completeness of protein sequences combined with a more accurate taxonomic assignment method is expected to yield a more reliable taxonomic annotation. From the 93 million proteins predicted by Augustus, 8,001,326 (present on 5,724,823 contigs from 4,083 metagenomes, Fig. 1) were assigned a eukaryotic taxonomic annotation by the Diamond LCA approach (Table 2) and are made available as a curated dataset of eukaryotic soil proteins[39] with taxonomic informations[40].

Of these 8 million proteins, 45% were assigned a Opisthokonta taxonomy (Fungi + Metazoa), of which 96% were fungal and only 4% Metazoa (Fig. 5). These proportions are consistent with eukaryotic taxonomic distribution previously described in the literature, reporting fungi as the most abundant eukaryotic microorganisms

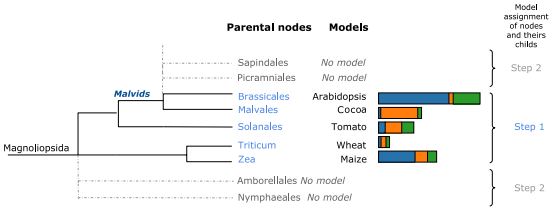


Fig. 3 Phylogenetic tree focused on Magnoliopsida clades displaying the Augustus model distribution supporting the assignment of *ab initio* gene model by dataset (blue = Plant-associated, orange = Terrestrial 1, green = Terrestrial 2).

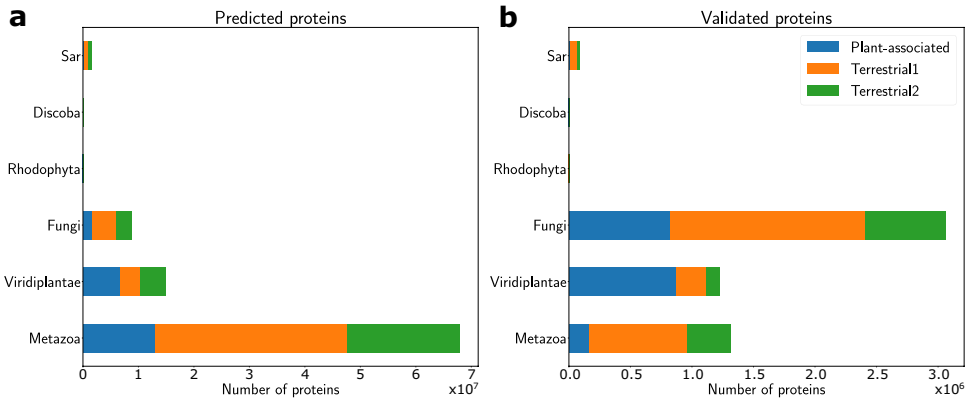


Fig. 4 Number of Augustus-predicted proteins and their taxonomic distribution per Augustus model kingdom by dataset (**a**) on all contigs (**b**) on eukaryotic contigs validated by Diamond (blue = Plant-associated, orange = Terrestrial 1, green = Terrestrial 2).

Clade	Plant-ass.	Terrestrial1	Terrestrial2	Total	%
Prokaryote	12,271,986	11,564,201	20,560,428	44,396,615	47.6
Eukaryote	4,986,024	1,951,235	1,064,070	8,001,326	8.6
Viruses	23,743	25,409	70,942	120,094	0.1
Undetermined	4,511,252	29,664,147	6,655,739	40,831,138	43.7
Total	21,793,005	43,204,992	28,351,179	93,349,176	100

Table 2. Taxonomic classification of Augustus predicted proteins in superkingdoms by the Last Common Ancestor algorithm of DIAMOND among each dataset.

in studied soil[2,41]. Actually, in soil metagenomes, fungal organisms are often second to bacteria in number and account for a comparable proportion of the biomass. Here, we retrieved 1,657 different fungal TaxIDs covering granularity levels ranging from species to the whole kingdom. Taxonomic annotations at deeper taxonomic nodes indicate the protein is equally related to proteins from multiple different and phylogenetically distinct fungal species. Among Metazoa, the dominant categories were Arthropoda, then Nematoda and Rotifera, respectively representing 48%, 9% and 8% of Metazoa (Supplementary Data[42]), again consistent with these species being the most abundant animals in soil environment. Besides Opisthokonta, Viridiplantae was actually the most represented kingdom, with 49% of all eukaryotic taxonomic assignment (Fig. 5). This suggests plant material is frequently present in soil samples and this is particularly expected for the plant-associated samples. Besides Opisthokonta and plants, other eukaryotes mainly belonged to the category SAR (1% of all) and most of the rest (5%) were unclassified eukaryotes (category other eukaryota, Fig. 5). These last taxa show small percentage of the whole dataset of soil eukaryotic proteins but still represent several thousand of proteins due to the size of the dataset.

The rest of the 93 million soil proteins were either assigned a non-eukaryotic TaxID with 47.6% and 0.1% being assigned a prokaryotic and viral taxonomy, respectively (Table 2), or had undetermined taxonomic annotation (43.7%).

Identification of potential orphan eukaryotic proteins. More than 40 million proteins, representing 43.7% of the total Augustus predictions, could not be assigned a prokaryotic, eukaryotic or viral TaxID. Among them 27,269,572 (67%) were assigned untraceable taxonomic identifiers such as 'unclassified' (e.g. '12908' TaxID) or 'other' (e.g. '32644' TaxID), and the rest of the proteins (13,561,566 or 33%) simply returned no hit at all against

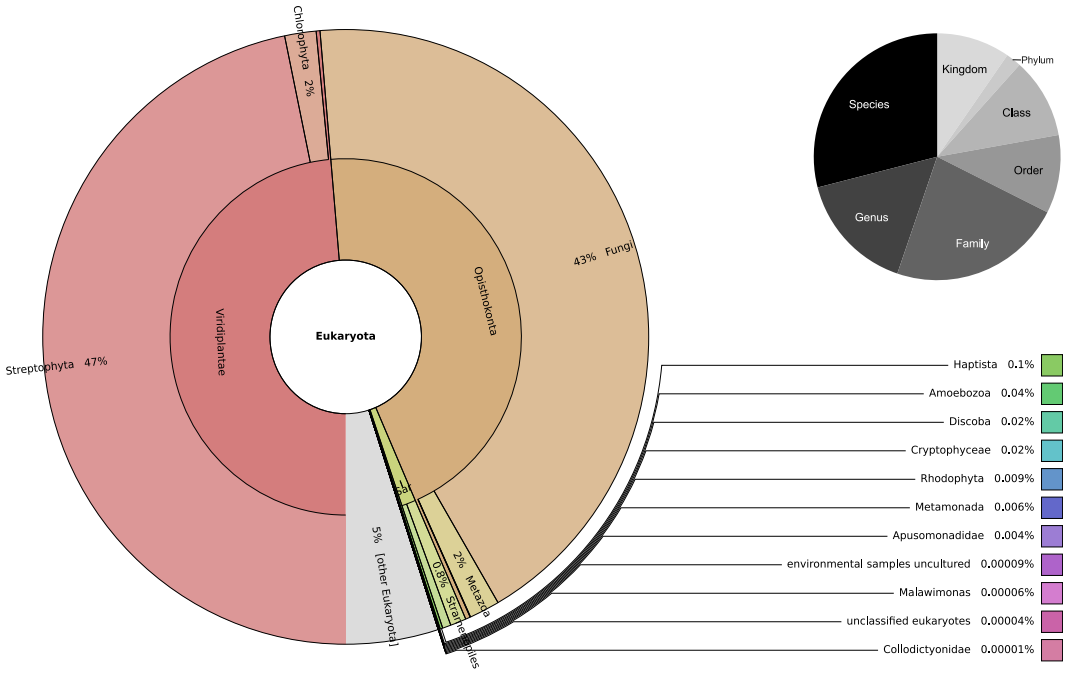


Fig. 5 Krona representation of taxonomic assignment provided by the last common ancestor algorithm of DIAMOND for the 8 million eukaryotic proteins predicted by our homemade pipeline using Augustus (HTML file: available on Supplementary Data[42]), and the pie chart of taxonomic ranks of retrieved lineages.

the NCBI's nr library in our Diamond homology search. Because these proteins might represent false positives from Augustus, they were not blindly added to the dataset of 8 million eukaryotic soil proteins. However, these proteins might as well represent orphan eukaryotic proteins lacking homology in public databases, constituting an important resource for new discoveries. To discriminate potential eukaryotic from non-eukaryotic orphan proteins, we assessed whether they were distributed on otherwise mostly-eukaryotic contigs. Thus, from the initial dataset of 113 million Kraken2-assessed eukaryotic contigs, we only retained orphan proteins present on contigs that contained at least 50% of Diamond-confirmed eukaryotic protein-coding genes. This yielded a total of 3,657,380 contigs distibuted on 4,059 metagenomes (Supplementary Data, Table 3[33]). A total of 354,243 orphan proteins were distributed on these contigs and represent potential novel eukaryotic proteins. We made this additional dataset of potential novel orphan eukaryotic proteins also available[43].

Reducing redundancy of soil eukaryotic proteins. Some redundancy was expected because we used metagenomic data from thousands of individual studies, and some sequencing data came from the same sampling location. Therefore, we clustered Fasta files using the Linclust software of the MMseq2 metagenomic toolkit[44]. For both eukaryotic and orphan datasets, we clustered proteins with at least 99% sequence identity and covering at least 90% of the target. With these parameters, the 8 million eukaryotic proteins were clustered in 4,624,994 representative sequences[45], and the 354,243 orphan proteins were clustered in 288,612 proteins[46]. For both clusterings, we provide the correspondence files to link original protein predictions to their respective representative clusters[47,48].

Data Records

All processed and Supplementary Data are publicly available on Data INRAE portal[49] containing files described in Table 3.

Technical Validation

Comparison of protein prediction and taxonomic annotation quality to original JGI annotation. To determine whether using Augustus in our pipeline allowed improving eukaryotic protein predictions, we compared them to the predicted proteins obtained by the JGI using Prodigal for the same set of contigs. For this comparison, we used the same 3,657,380 contigs (covering 4,059 different metagenomes) containing at least 50% of predicted proteins with a eukaryotic taxonomy assigned by Diamond-LCA (defined above). Our pipeline allowed predicting 5.6 million proteins in these contigs. In comparison, on the same dataset, Prodigal initially predicted a total of 16 million proteins, covering 3,294,764 of these contigs and 3,979 metagenomes (Supplementary Data, Table 3[33]). First, although the number of protein predicted is higher with Prodigal, this software was unable to predict proteins in more than 360,000 contigs (3,657,380-3,294,764). Moreover, the raw number of proteins can be misleading because while Prodigal predicted 1,9 billion amino acids, our methodology allowed predicting 2.5 billion amino acids in total, suggesting although more proteins were predicted by Prodigal, they were much shorter and probably fragmented. Augustus allowed predicting introns in 1,627,033 genes from 1,074,415 contigs; these intronic sequences span on average 17% of the gene length. In comparison, Prodigal is

File name	Type	Size	Path	Description
eukaryotic_proteins.aa[39]	fasta	3GB	.	8 M of validated eukaryotic proteins predicted with Augustus in contigs from Terrestrial and Plant-associated metagenomic data from JGI
eukaryotic_proteins_taxonomy.txt[40]	text file	1,9GB	.	Taxonomic information for 8 M of validated eukaryotic proteins from the last common ancestor algorithm of Diamond
orphan_Euka.aa[43]	fasta	79MB	.	Orphan proteins from contigs with over half of eukaryotic proteins
eukaryotic_proteins_clustered.aa[45]	fasta	1.8GB	.	4,6 M representative clusters of 8 M of eukaryotic proteins
eukaryotic_proteins_clustered.tsv[47]	TSV	614MB	.	Composition of eukaryotic protein clusters
orphan_proteins_clustered.aa[46]	fasta	66MB	.	288,612 representative clusters of orphan proteins
orphan_proteins_clustered.tsv[48]	TSV	27MB	.	Composition of orphan protein clusters
eukaryotic_proteins_taxonomy_krona.html[42]	html	1,7MB	./Supplementary Data	Krona representation of 8 M of validated eukaryotic protein taxonomy from last common ancestor algorithm of Diamond
Supplementary_data_1.txt[24]	text file	158KB	./Supplementary Data	List of metagenome identifier of processed data from JGI
Supplementary_data_Figures.pdf[32]	PDF	323KB	./Supplementary Data	Fig. 1: Informations on eukaryotic proteins prediction processing
				Fig. 2: BUSCO scores by dataset
Supplementary_data_tables.pdf[33]	PDF	51KB	./Supplementary Data	Table 1: Kraken2 lineage distribution in main eukaryotic Clade
				Table 2: Number of proteins predicted with Augustus
				Table 3: Information on gene prediction outputs
				Table 4: Statistics of BUSCO scores

Table 3. Data record, information about files available on public repository DATA INRAE[49].

		BUSCO scores					Fasta informations		
	Model	Complete	Complete Single	Complete Duplicated	Fragmented	Missing	Nb. of Proteins	Total nb. of AA	Nb. of AA/ protein
1	Mix	100	12.9	87.1	0	0	63,986	25,941,958	405
2	Fusarium	98.4	12.5	85.9	1.2	0.4	87,508	36,614,755	418
3	Zebrafish	96.1	23.9	72.2	3.1	0.8	152,796	43,294,314	283
4	Metaeuk nr	100	1.6	98.4	0	0	119,085	34,031,250	286
5	MetaEuk swp	97.6	8.2	89.4	0.8	1.6	34,906	12,112,481	347
6	Prodigal	77.3	36.9	40.4	20	2.7	271,456	37,520,032	138

Table 4. BUSCO scores and FASTA files information for several gene prediction methods (1) Augustus with a mixture of model as in our paper, (2) Augustus with Fusarium model, (3) Augustus with Zebrafish model, (4) MetaEuk with NR database, (5) MetaEuk with Swissprot database and (6) Prodigal. All scores are computed on the same metagenome used as reference.

not able to predict introns and ends its prediction when the first stop codon is encountered. Therefore, at least 28% (1.6/5.6 millions) of the proteins predicted by Augustus were necessarily incorrectly predicted by Prodigal, initially. Moreover with the high frequency of stop codons in the intronic regions due to less selective pressure on these genomic regions, most intron-containing genes are expected to be truncated by Prodigal. Overall, we oberve that our strategy was able to predict longer proteins and on more contigs that the initial Prodigal annotation. Hence, to further compare predictions from both methods, we used two metrics: *(i)* protein length distribution, and *(ii)* the recovery of nearly universal single copy eukaryotic genes.

Length distribution of protein. First, we calculated and compared the distribution of protein lengths from Augustus vs. Prodigal predictions. Proteins predicted by Augustus were significantly longer than proteins predicted by Prodigal on the same contigs (Fig. 6; unpaired t-test, $n = 5.3\ 10^6/n = 9\ 10^6$ proteins, $T = 1.994\ 10^3$, $p \leq 10^{-4}$). These observations coupled with the higher number of proteins predicted by Prodigal, confirm that Augustus was able to predict introns and join together multiple exons to form more complete genes where Prodigal predicted multiple truncated genes. Of note, the average size of genes (in ext. proteins) in eukaryotes is larger than in prokaryotes, due to the evolution of genome complexity[50]. Furthermore, the length distribution is closer to a normal one with Augustus predictions than with Prodigal ones (Fig. 6b), indicating a better quality of our new predictions. Indeed, Nevers *et al.*[51] reports that a non-normal distribution of proteins length, as observed for these Prodigal predictions in eukaryotic contigs, is indicative of more truncated proteins caused by fragmented genomes and incorrect protein prediction. Overall, the authors showed that protein lengths

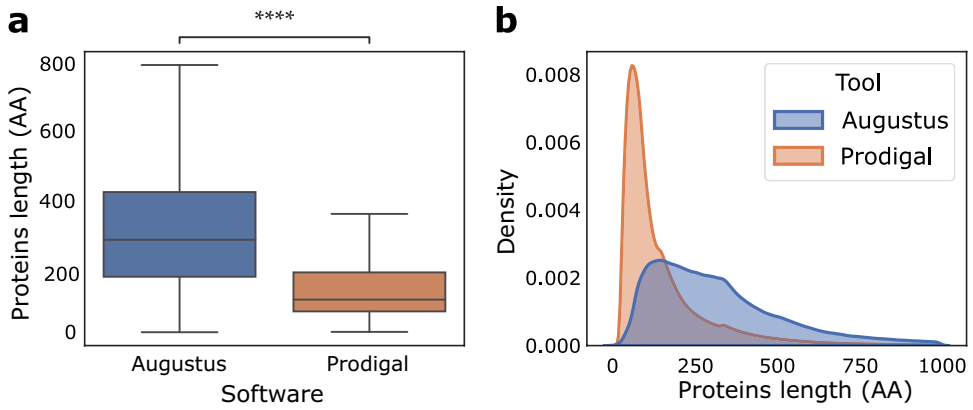


Fig. 6 Distribution of protein lengths of Augustus prediction in blue versus Prodigal prediction in orange. Proteins from Augustus are significantly longer than those from Prodigal (see text).

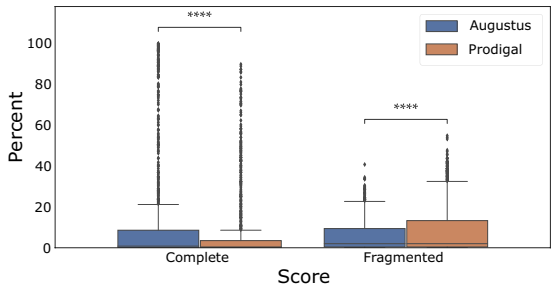


Fig. 7 Complete and Fragmented BUSCO scores of the 1,093 metagenomes with single-copy universally conserved genes report a significantly better recovery of genes from eukaryotic microorganisms with Augustus than Prodigal (see text).

distribution is remarkably well conserved across species and this feature could be used as quality metric in addition to other measures.

Recovery of nearly universal single-copy eukaryotic genes. To assess the improvement of our *de novo* eukaryotic protein predictions from soil microorganisms, we also compared the proportions of near-universal single-copy orthologs retrieved for each metagenome with those provided by Prodigal in the same contigs using BUSCO (v.4.0.2) in protein mode with '*eukaryota_odb10*' lineage[52]. Starting from the 4,059 metagenomes containing contigs with at least 50% eukaryotic proteins, universally-conserved eukaryotic BUSCO proteins were identified in contigs coming from 1,093 metagenomes. This observation is not particularly surprising since *(i)* there are only 255 universally-conserved eukaryotic BUSCO genes, *(ii)* eukaryotes represent a minority of species in the soil[2] and *(iii)* most eukaryotic genomes are only partially assembled from short-read based on shotgun metagenomic data.

The proportion of BUSCO genes found in complete length in metagenomes was significantly higher for the Augustus predictions than for the initial Prodigal predictions (Fig. 7; paired Wilcoxon-test, n = 1,093 metagenomes, Complete $T = 1.132\ 10^5$, $p \leq 10^{-4}$; Fragmented $T = 2.039\ 10^4$, $p \leq 10^{-4}$). Similarly, the proportion of fragmented and missing BUSCO genes were significantly lower in Augustus predictions as compared to Prodigal predictions; this trend is identical for all datasets (Supplementary Data, Fig. 2[32]). BUSCO completeness scores from our Augustus gene predictions are as good or better than Prodigal for more than 98% of metagenomes. Furthermore, we have predicted more universal single-copy genes than Prodigal for 574 metagenomes, or more than half of the 1,093 metagenomes containing at least one BUSCO gene in one of both predictions. We observe an average improvement of 11.9% in the BUSCO completeness score, and genes are less fragmented in 510 metagenomes with an average of 8.5% lower proportion of fragments (Supplementary Data, Table 4[33]). The scores provided by BUSCO for these 1,093 metagenomes show a significant improvement of protein recovery and completeness for proteins from our Augustus-based strategy as compared to those from Prodigal, indicating our pipeline has improved the quality of eukaryotic gene models in soil metagenomes.

Accuracy and diversity of taxonomic annotation. We assessed whether the Diamond-LCA taxonomic annotation strategy we employed allowed gaining information over the original JGI taxonomic annotation. To perform this evaluation, we compared the richness of taxonomic information proposed by our strategy to the original JGI annotation on a group of eukaryotic soil microorganisms known to play important ecological roles, Arbuscular

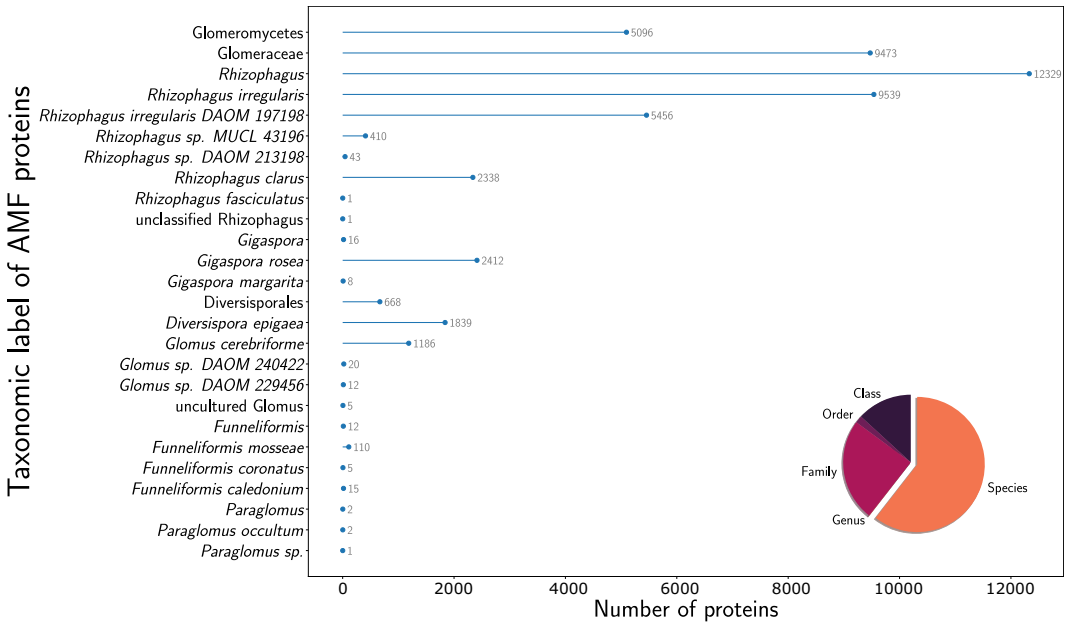


Fig. 8 Annotated taxa of Arbuscular Mycorrhizal Fungal proteins with the last common ancestor algorithm of Diamond after protein prediction with Augustus. Number of proteins is shown for each taxa. The ratio of the taxonomic rank of annotations across AMF lineages is shown in a pie chart.

Mycorrhizal Fungi (AMF). Indeed, AMF are ubiquitous members of soil microbiota, and more particularly of the (plant-associated) rhizosphere[15]. These eukaryotic microorganisms are plant symbionts with high impacts in several fields, mainly in agronomy due to a bio-stimulant and a bio-protective effect[16,53], but they are also used to help in environmental issues such as cleaning-up polluted soils or facilitating reforestation. Among all the contigs containing at least 50% of eukaryotic proteins, according to Diamond-LCA, only 8,065 AMF proteins were predicted in the original JGI annotation, covering 6,048 contigs from 327 metagenomes. Moreover, all these proteins were assigned the same and sole AMF species/TaxID: *Rhizophagus irregularis*. In contrast, using our eukaryotic-centred gene prediction and taxonomic annotation pipeline, we expand the identification of AMF to 50,999 proteins in 48,726 contigs from 1,102 metagenomes. Furthermore, this new annotation now covers 26 different taxa (from class to species) better representing the AMF diversity present in these soils (Fig. 8). The case of these pervasive eukaryotic microorganisms in the soil highlights the benefits of this work to improve the representation of eukaryotic organisms in public soil metagenomes[40].

Validation of taxonomic assignment and gene prediction strategy. *Taxonomic assignment methods.* Comparing the Kraken2-assigned Augustus models for gene prediction on contigs to the taxonomic assignment at the protein-level based on Diamond LCA, we observed substantial differences in the relative proportion of taxonomic groups (Figs. 4, 5). For instance, while metazoan Augustus models were assigned to 49 and 76% of contigs in plant-associated and terrestrial datasets, respectively, only 2% of the eukaryotic proteins were assigned a metazaon taxonomy via Diamond-LCA. Conversely, while fungal Augustus models were assigned to only 6.5% of contigs, fungi represented 43% of taxonomic assignments obtained by Diamond-LCA. These Diamond-LCA taxonomic assignments are more consistent with the expected dominant taxa in the soil and illustrate the interest of our two-steps strategy with the first sensitive step aiming at identifying as many putative eukaryotic contigs as possible and the second specific step aiming at assigning an as reliable as possible taxonomic annotation to the genes and proteins. Furthermore at a global level, of the 93 million proteins predicted on the contigs deemed eukaryotic according to Kraken2, only 8 millions could be confirmed as eukaryotic with Diamond-LCA. An explanation for this discrepancy between Kraken2 and Diamond-LCA taxonomic assignments may be the following. A substantial proportion of contigs were probably assigned a eukaryotic taxonomy by Kraken2 based on a low number of k-mer matching with the eukaryotic target. The proteins predicted on these contigs were not assigned a eukaryotic annotation by Diamond but either a prokaryotic or undetermined taxonomy. Applying a confidence score threshold to Kraken2 taxonomic predictions might have resolved part of these false positives but at the risk of augmenting the rate of false negatives, and thus missing many eukaryotic contigs. Because we wanted this first filtering step to be as sensitive as possible, we decided not to apply a stringent confidence test on Kraken2 and to rely on further Diamond-based LCA strategy for more accurate final taxonomic annotation at the protein level.

We also compared the original single best BLAST hit JGI strategy for taxonomic annotation to the Diamond-LCA taxonomic assignment we employed in this study. Using a single best BLAST hit strategy, all taxonomic annotations were necessarily at the species level, regardless of the other hits and regardless of the percent identity with the best hit. This strategy can be misleading, in particular if the taxonomic annotation of the best hit is erroneous or if the similarity is only distant and to a variety of different species with no jump in E-values.

In contrast, using an LCA approach, we noticed that only less than 30% of the proteins are still annotated at the species level. This indicates the rest of the proteins have been assigned a deeper taxonomic rank (Fig. 5) because they matched multiple hits with similarly good scores. This re-assignment of taxonomic annotation to deeper, more ancestral level decreases the risk of making errors by assigning a very shallow and precise taxonomic annotation based on spurious or distantly related best BLAST hits. This situation is expected to be particularly frequent when annotating proteins from environmental samples returning only distant similarity to proteins present in reference protein libraries from cultured organisms[35].

Assessment of gene prediction strategy. As mentioned in previous sections, taxonomic annotations at nucleic (contig) and protein scales are not necessarily consistent. This fact may cast doubt on Augustus model selection procedure. To evaluate our soil eukaryotic gene prediction strategy, we compared the quality of proteins obtained by a mixture of Augustus models selected by our pipeline with prediction using either only one Augustus model, chosen as (i) Fusarium (a fungal model retrospectively corresponding to the most represented taxon in the final Diamond-based taxonomic assignment), or (ii) Zebrafish (a metazoan model corresponding to a taxon with low chance to be actually present in soil contigs), or (iii) another gene prediction software, MetaEuk[54] with NR database as a reference and finally (iv) MetaEuk with SwissProt database as a reference. These four strategies were used to predict proteins on Kraken2-assigned eukaryotic contigs within the same dataset: metagenome '3300031471' from the terrestrial dataset. This metagenome was randomly chosen among those containing complete eukaryotic BUSCO genes and thus representing an easy reference to check whether our mutli-model Augustus approach was relevant compared to single-model or reference database approaches. We compared eukaryotic BUSCO scores as well as the number of predicted proteins and the number of amino-acids per protein. Concerning Augustus, we observed the best recovery of universally conserved genes using our procedure (mixture of phylogenetically assigned models) (Table 4; line 1,2,3). Hence, although fungi and in particular *Fusarium* were the most numerous taxa in soil metagenomes, a mixture of models chosen by our procedure allowed a better recovery of BUSCO proteins. Thus, despite a necessarily substantial portion of imperfect model assignments, due to discrepancy between *a priori* Kraken2 taxonomic assignment and *a posteriori* Diamond taxonomic confirmation, a mixture of models seem to yield better results than a single phylogenetically close model. This is probably due to complex nature of soil communities. In contrast, and as expected, assigning a fish model for this soil sample returned the lowest BUSCO completeness and the highest proportion of fragmented and missing proteins. Concerning MetaEuk, BUSCO results were as good as our mixture of models procedure, when using the NCBI's nr library as a reference (Table 4; line 1,4,5). However, a comparison of protein lengths distribution suggested that, besides BUSCO proteins, MetaEuk protein predictions were globally shorter with more proteins, a lower average number of amino acids per protein and a lower median length (Table 4). We tried whether changing the reference library in Metaeuk would improve protein lengths distribution by using Swissprot instead of nr. Using Swissprot indeed improved protein length metrics although these metrics were not as good as for our procedure, and came at the cost of decreased BUSCO completeness (Table 4, line 5). Overall, it seems that Metaeuk is more sensitive than the multi Augustus model we selected as more proteins were predicted. However, these proteins are shorter and might either represent short actual proteins or fragments. Our strategy was to be permissive at the contig level but stringent at the protein level (e.g. not to search for proteic 'dark matter'). Although erroneous annotations inherent to massive high-throughput *de novo* gene prediction approaches can remain on some eukaryotic contigs, using Augustus with a mixture of gene models seems to represent the optimal balance between recovery of complete BUSCO genes and prediction of the longest and less fragmented proteins besides BUSCO ones.

Usage Notes

Current microbiology investigations are focused on addressing the factors shaping the structure of microbial communities. To drive the development of tomorrow's biotechnology it is essential to understand biological pathways both at the organism level and at the inter-microbial relationships scale, for prokaryotic and eukaryotic organisms together. This dataset provides a more complete and comprehensive view of the pool of genes and proteins, genetic diversity and distribution of eukaryotic microbes in soil and plant-associated microbiomes. At the molecular level, the use of this data is relevant to address biological questions in both fundamental research on plant-microbe interactions and applied, agronomical research, such as the study of potential metabolic functions of telluric eukaryotes, or of the interaction pathways between microbial members of the community. At a broader scale, the more accurate taxonomic annotation provides an unparalleled opportunity to assess how microbial eukaryotes are distributed across the soil and plant-associated microbial-environments. As illustrated in the data validation section, this improvement of microbial eukaryote representation has allowed us to increase by a factor of six times the detection of the ubiquitous AMF species, which are of high agronomic and economic interest.

Any research involving study of soil eukaryotes from evolutionary research on gene flow and transfers within the biome to more translational research aiming at deciphering important soil functions and biochemical pathways will benefit from this improved dataset of soil proteins with more accurate taxonomic annotation. In addition, our data can be cross-referenced with the metadata provided by the JGI (downloadable from the IMG/M portal) which includes geo-tracking and a wealth of environmental, sampling and processing information on each metagenome. They can be linked to proteins and annotations by searching for the 'metagenomeID', as each protein name in our dataset has a nomenclature based on the following pattern: 'contigName_metagenomeID.geneID', to offer this possibility. On one hand, the ecological metadata provides an unprecedented potential to study the effect of the environment on community structures and to have a better, more comprehensive view on how external factors influence the eukaryotic soil microbial communities. On the other hand, metadata on sampling and processing could be useful to assess which parameters affect the diversity and sequencing of

eukaryotes in metagenomes and help to shape future protocols. Moreover, in this study, we provided a fully documented pipeline and protocol available as python scripts from the detection of putative eukaryotic contigs to the *ab initio* model selection for Augustus gene prediction and further Diamond-based taxonomic annotation, that can be re-used to improve the annotation of eukaryotes on any microbiome data, including in other biomes than the soil.

Code availability

- Project name: EukaProt_in_PublicSoilMetag[31]
- Project home page: https://github.com/CaroleBelliardo/EukaProt_in_PublicSoilMetag.git
- Operating system(s): Platform independent
- Programming language: Python3
- Other requirements: Python3.8 or higher
- License: License: GNU General Public License v3.0

Received: 24 November 2021; Accepted: 26 May 2022;
Published online: 16 June 2022

References

1. Thiele-Bruhn, S. The role of soils in provision of genetic, medicinal and biochemical resources. *Philosophical Transactions of the Royal Society B: Biological Sciences* **376**, 20200183, https://doi.org/10.1098/rstb.2020.0183 (2021).
2. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* **15**, 579–590, https://doi.org/10.1038/nrmicro.2017.87 (2017).
3. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding: NEXT-GENERATION DNA METABARCODING. *Molecular Ecology* **21**, 2045–2050, https://doi.org/10.1111/j.1365-294X.2012.05470.x (2012).
4. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in new york city's central park are similar to those observed globally. *Proc. R. Soc. B.* **281**, 20141988, https://doi.org/10.1098/rspb.2014.1988 (2014).
5. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nature Biotechnology* **39**, 499–509, https://doi.org/10.1038/s41587-020-0718-6 (2021).
6. Naylor, D. *et al.* Deconstructing the Soil Microbiome into Reduced-Complexity Functional Modules. *mBio* **11**, https://doi.org/10.1128/mBio.01349-20 (2020).
7. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533–1542, https://doi.org/10.1038/s41564-017-0012-7 (2017).
8. Bach, E. M., Williams, R. J., Hargreaves, S. K., Yang, F. & Hofmockel, K. S. Greatest soil microbial diversity found in micro-habitats. *Soil Biology and Biochemistry* **118**, 217–226, https://doi.org/10.1016/j.soilbio.2017.12.018 (2018).
9. Dupont, A. O. C., Griffiths, R. I., Bell, T. & Bass, D. Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs: Soil pH and protistan diversity. *Environ Microbiol* **18**, 2010–2024, https://doi.org/10.1111/1462-2920.13220 (2016).
10. Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science* **346**, 1256688, https://doi.org/10.1126/science.1256688 (2014).
11. Torsvik, V. & Øvreås, L. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology* **5**, 240–245, https://doi.org/10.1016/S1369-5274(02)00324-7 (2002).
12. Jansson, J. K. & Hofmockel, K. S. Soil microbiomes and climate change. *Nature Reviews Microbiology* **18**, 35–46, https://doi.org/10.1038/s41579-019-0265-7 (2020).
13. Bonkowski, M. Protozoa and plant growth: the microbial loop in soil revisited. *New Phytologist* **162**, 617–631, https://doi.org/10.1111/j.1469-8137.2004.01066.x (2004).
14. Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. & Hay, S. I. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**, 214–217, https://doi.org/10.1038/nature03342 (2005).
15. Bonfante, P. & Genre, A. Plants and arbuscular mycorrhizal fungi: an evolutionary-developmental perspective. *Trends in Plant Science* **13**, 492–498, https://doi.org/10.1016/j.tplants.2008.07.001 (2008).
16. Schouteden, N., De Waele, D., Panis, B. & Vos, C. M. Arbuscular mycorrhizal fungi for the biocontrol of plant-parasitic nematodes: A review of the mechanisms involved. *Front. Microbiol.* **6**, https://doi.org/10.3389/fmicb.2015.01280 (2015).
17. Tran, B. T. T., Watts-Williams, S. J. & Cavagnaro, T. R. Impact of an arbuscular mycorrhizal fungus on the growth and nutrition of fifteen crop and pasture plant species. *Functional Plant Biology* **46**, 732, https://doi.org/10.1071/FP18327 (2019).
18. Bonfim, J. A. *et al.* Diversity of Arbuscular Mycorrhizal Fungi in a Brazilian Atlantic Forest Toposequence. *Microbial Ecology* **71**, 164–177, https://doi.org/10.1007/s00248-015-0661-0 (2016).
19. Hao, Z., Xie, W. & Chen, B. Arbuscular Mycorrhizal Symbiosis Affects Plant Immunity to Viral Infection and Accumulation. *Viruses* **11**, 534, https://doi.org/10.3390/v11060534 (2019).
20. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes fromcomplex natural microbial communities. *Genome Res.* **28**, 569–580, https://doi.org/10.1101/gr.228429.117 (2018).
21. Lind, A. L. & Pollard, K. S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9**, 58, https://doi.org/10.1186/s40168-021-01015-y (2021).
22. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* **45**, D507–D516, https://doi.org/10.1093/nar/gkw929 (2017).
23. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, https://doi.org/10.1186/1471-2105-11-119 (2010).
24. Belliardo, C. *et al.* metagenomeid.txt. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/N0HHAK (2021).
25. Murray, B., Dailey, M., Ertekin, E. & DiRuggiero, J. Draft metagenomes of endolithic cyanobacteria and cohabitants from hyper-arid deserts. *Microbiol Resour Announc* **10**, e0020621, https://doi.org/10.1128/MRA.00206-21 (2021).
26. Ward, R. D. *et al.* Metagenome sequencing to explore phylogenomics of terrestrial cyanobacteria. *Microbiol Resour Announc* **10**, https://doi.org/10.1128/MRA.00258-21 (2021).
27. Brent, M. R. How does eukaryotic gene prediction work? *Nat Biotechnol* **25**, 883–885, https://doi.org/10.1038/nbt0807-883 (2007).
28. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257, https://doi.org/10.1186/s13059-019-1891-0 (2019).
29. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745, https://doi.org/10.1093/nar/gkv1189 (2016).

30. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
31. Belliardo, C. CaroleBelliardo/EukaProt_in_publicsoilmetag. *Zenodo* https://doi.org/10.5281/ZENODO.6546146 (2022).
32. Belliardo, C. *et al*. Supplementary_data_figures.pdf. In *Improvement of eukaryotic protein predictions from soil metagenomes*, https://doi.org/10.15454/E2VTRB/GAKY0C (Portail Data INRAE, 2021).
33. Belliardo, C. *et al*. Supplementary_data_tables.pdf. In *Improvement of eukaryotic protein predictions from soil metagenomes*, https://doi.org/10.15454/E2VTRB/Y6L2OH (Portail Data INRAE, 2021).
34. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
35. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927, https://doi.org/10.1093/bioinformatics/btz848 (2020).
36. Cribdon, B., Ware, R., Smith, O., Gaffney, V. & Allaby, R. G. PIA: More accurate taxonomic assignment of metagenomic data demonstrated on sedaDNA from the north sea. *Front. Ecol. Evol.* **8**, 84, https://doi.org/10.3389/fevo.2020.00084 (2020).
37. Eisenhofer, R. & Weyrich, L. S. Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ* **7**, e6594, https://doi.org/10.7717/peerj.6594 (2019).
38. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Research* **17**, 377–386, https://doi.org/10.1101/gr.5969107 (2007).
39. Belliardo, C. *et al*. eukaryotic_proteins.aa. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/T1OHIX (2021).
40. Belliardo, C. *et al*. eukaryotic_proteins_taxonomy.txt. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/A1TUGT (2021).
41. Lesaulnier, C. *et al*. Elevated atmospheric CO_2 affects soil microbial diversity associated with trembling aspen. *Environ Microbiol* **10**, 926–941, https://doi.org/10.1111/j.1462-2920.2007.01512.x (2008).
42. Belliardo, C. *et al*. eukaryotic_proteins_taxonomy_krona.html. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/A2BOIB (2021).
43. Belliardo, C. *et al*. orphan_euka.aa. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/3XPVTN (2021).
44. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542, https://doi.org/10.1038/s41467-018-04964-5 (2018).
45. Belliardo, C. *et al*. eukaryotic_proteins_clustered.aa. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/1TK3RE (2021).
46. Belliardo, C. *et al*. orphan_proteins_clustered.aa. *In* Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/NO0Z7D (2021).
47. Belliardo, C. *et al*. eukaryotic_proteins_clustered.tsv. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/TFJJKX (2021).
48. Belliardo, C. *et al*. orphan_proteins_clustered.tab. In Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB/54EDIJ (2021).
49. Belliardo, C. *et al*. Improvement of eukaryotic protein predictions from soil metagenomes. *Portail Data INRAE* https://doi.org/10.15454/E2VTRB (2021).
50. Xu, L. *et al*. Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. *Molecular Biology and Evolution* **23**, 1107–1108, https://doi.org/10.1093/molbev/msk019 (2006).
51. Nevers, Y., Defosset, A. & Lecompte, O. Orthology: Promises and challenges. In Pontarotti, P. (ed.) *Evolutionary Biology—A Transdisciplinary Approach*, 203–228, https://doi.org/10.1007/978-3-030-57246-4_9 (Springer International Publishing, 2020).
52. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* msab199, https://doi.org/10.1093/molbev/msab199 (2021).
53. Hoysted, G. A. *et al*. A mycorrhizal revolution. *Current Opinion in Plant Biology* **44**, 1–6, https://doi.org/10.1016/j.pbi.2017.12.004 (2018).
54. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48, https://doi.org/10.1186/s40168-020-00808-x (2020).

Acknowledgements

We would like to warmly thank for they help and support all members of the bioinformatics platform of the Institute Sophia Agrobiotech, Sophia Antipolis, France. We also thank for his advice on soil microbial analyses Samuel Mondy from INRAE, Dijon, France. Finally, we would like to thank all the persons who worked to generate the data publicly available on the IMG/M platform. In particular, we thank the following labs for having granted us the right to use their data despite the use-restrictions: Dr. Rich V. and the NSF Biology EMERGE Integration Institute, (NSF-BII 2022070); Dr. DiRuggiero J. and The Johns Hopkins Department of Biology; Dr. Pietrasiak and NMSU Plant and Environmental Sciences department of College of Agricultural, Consumer, and Environmental Sciences (ACES). *Raw data were produced by the US Department of Energy Joint Genome Institute (*https://ror.org/04xm1d337*; operated under Contract No. DE-AC02-05CH11231) in collaboration with the user community*.

Author contributions

C.B. co-designed the study, implemented programs, compiled and validated data, wrote and revised the manuscript, G.D.K. conceived part of the experiments, C.R. conceived part of the experiments, M.C. reviewed and validated the paper, acquired funding J.L. reviewed and validated the paper, acquired funding M.B.-B. co-designed the study, participated in data analysis, figure preparation and manuscript revision, acquired funding E.G.J.D. co-designed the study, participated in data analysis, wrote and revised the manuscript, acquired funding. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.B. or E.G.J.D.

3. *Conclusion et perspectives*

L'ensemble de ce travail a permis d'optimiser la représentation des micro-organismes eucaryotes dans les métagénomes de sols en améliorant la qualité des protéines prédites et leur classification taxonomique, comme illustré avec l'exemple des protéines issues de champignons mycorhiziens à arbuscules dans l'article (Belliardo et al., 2022). En effet, seulement 8 065 protéines avaient été identifiées comme provenant de CMA dans 327 métagénomes dans les annotations provenant du JGI, et l'ensemble de ses protéines avait été assigné à la même espèce '*Rhizophagus irregularis*'. Dans notre travail, l'abondance et la diversité des protéines de CMA ont été significativement enrichies avec maintenant 50 990 protéines prédites dans 1 102 métagénomes différents et des assignations taxonomiques correspondant à une vingtaine de lignées de CMA.

Le jeu de données résultat de ce travail a été mis à disposition de la communauté scientifique dans « Recherche Data Gouv » et pourra être utilisé dans le cadre de recherches portant sur les microbiomes terrestres. Pour favoriser la diffusion de ces données, un communiqué scientifique intitulé « Amélioration de la prédiction des protéines eucaryotes pour mieux comprendre l'écologie et l'évolution des micro-organismes du sol en interactions avec les plantes » rédigé avec M. Arnaud Ridel, a été publié par le département SPE. Ce travail a aussi été proposé comme fait marquant 2022 de notre unité.

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

AGROÉCOLOGIE 3 min

Amélioration de la prédiction des protéines eucaryotes pour mieux comprendre l'écologie et l'évolution des microorganismes du sol en interactions avec les plantes

Des chercheurs INRAE de l'Institut Sophia Agrobiotech en collaboration avec la société MYCOPHYTO ont extrait et réanalysé 6 800 jeux de données métagénomiques de sols afin d'améliorer la représentation de la biodiversité des micro-organismes eucaryotes présents dans les sols, dont les champignons mycorhiziens utilisés en tant que biostimulants.

Publié le 30 juin 2022

Une collaboration INRAE- MYCOPHYTO

Des chercheurs INRAE de l'Institut Sophia Agrobiotech en collaboration avec la société MYCOPHYTO viennent de publier dans la revue Scientific Data leurs travaux de recherche sur l'amélioration de la prédiction des protéines présentes dans les génomes des micro-organismes eucaryotes vivant dans les sols. Ce travail d'analyse a été réalisé à partir de données de séquençage de milliers d'échantillons de sols disponibles sur des serveurs de données publics. Cela a permis d'obtenir une meilleure représentation des eucaryotes composant les communautés microbiennes terrestres mais aussi de révéler la diversité des protéines codées par leurs génomes.

Ces travaux de recherche ont été réalisés dans le cadre de la thèse de Carole Belliardo, bio-informaticienne dans l'équipe Génomique et Evolution Moléculaire Adaptative (GAME) de l'Institut Sophia Agrobiotech. Après un Master « Science de la Vie et de la Santé » à l'Université de Nice, Carole Belliardo a obtenu en 2019 un co-financement de thèse par le département Santé des Plantes et Environnement d'INRAE et la société MYCOPHYTO qui propose aux agriculteurs des biostimulants basés sur la synergie entre les champignons mycorhiziens présents dans les sols et les racines des plantes.

Ces recherches ont un double intérêt pour les collaborateurs. Pour INRAE, ce travail est essentiel en termes de recherche fondamentale car il améliore nettement la qualité des séquences des protéines eucaryotes présentes dans le sol et donne une meilleure représentation de leur biodiversité. Pour la société MYCOPHYTO, l'intérêt se situe au niveau de l'exploitation de ces données qui permet d'extraire des informations concernant la distribution géographique des différentes espèces de champignons mycorhiziens dans les sols.

Améliorer la représentation de la diversité des micro-organismes des sols

Pour mener à bien ses recherches, Carole Belliardo a utilisé des données métagénomiques. L'approche métagénomique provient du séquençage sans distinction de l'ensemble des micro-organismes vivants dans un environnement spécifique. Pour cela, elle a utilisé les données stockées sur le serveur public IMG/M du Joint Genome Institute aux Etats-Unis. Elle a ainsi fait l'acquisition de plus de 6 800 jeux de données métagénomiques correspondant à des échantillons terrestres et de rhizosphères (ce qui représente plusieurs dizaines de téraoctets de données).

Cependant, la doctorante s'est rapidement aperçue que les protéines de micro-organismes eucaryotes prédites dans ces métagénomes étaient mal représentées et mal annotées dans ces données. Cette limitation est due aux procédures d'analyses proposées par les serveurs publics qui utilisent des outils dédiés au traitement de données procaryotes (bactéries et archées). Les micro-organismes eucaryotes sont ainsi négligés bien qu'ils représentent une biomasse souvent équivalente à celle des procaryotes dans les sols. Il était donc essentiel pour elle et l'ensemble de l'équipe de recherche d'améliorer la représentation des microorganismes eucaryotes dans ces données métagénomiques.

Elle a ainsi développé au cours de sa thèse un pipeline bio-informatique qui permet de détecter les séquences d'ADN provenant de micro-organismes eucaryotes dans des données métagénomiques et d'effectuer une prédiction de novo des protéines à l'aide d'un logiciel prenant en considération les caractéristiques génomiques des eucaryotes. Le déploiement de cette méthode sur les 7,9 milliards de génomes ou fragments de génomes microbiens du JGI a permis d'identifier 8 millions de protéines provenant d'organismes eucaryotes vivants dans les sols, et plus de 300 000 protéines orphelines. Ces protéines orphelines annotées dans des génomes ou fragments de génomes eucaryotes et dépourvues d'homologie dans les bases de données publiques représentent un potentiel de découverte sans précédent. Outre la caractérisation du potentiel fonctionnel déduit des analyses protéiques, il est possible d'assigner ces protéines à une information taxonomique permettant ainsi de coupler diversité taxonomique et fonctionnelle. Ces nouvelles données qui ont ainsi pu être générées, sont aujourd'hui partagées et accessibles sur un serveur public afin d'être réutilisées.

Choisir le bon "mix" de champignons mycorhiziens comme biostimulant

La société MYCOPHYTO, co-financeur de cette thèse, propose aux agriculteurs des champignons mycorhiziens comme biostimulants pour leurs cultures. L'objectif de MYCOPHYTO, avec ce projet de recherche, est de développer un algorithme qui permet d'identifier quel est le bon "mix" d'espèces de champignons mycorhiziens à utiliser en fonction des cultures et des paramètres physico-chimiques de l'environnement. Il s'agit d'optimiser le produit proposé à l'agriculteur.

Ainsi, le travail de Carole Belliardo est d'un grand intérêt pour MYCOPHYTO avec l'extraction des métadonnées associées aux séquences biologiques telles que la géolocalisation ou d'autres paramètres environnementaux. Cela permet de récupérer notamment des informations prépondérantes de distribution géographique des champignons mycorhiziens. Ces informations de distribution couplées avec des données de paramètres physico-chimiques de sols, météorologiques ainsi que d'autres données environnementales permettraient de développer un outil de prédiction fiable afin de déterminer le meilleur "mix" de champignons mycorhiziens à utiliser en fonction des paramètres environnementaux des sols.

RÉFÉRENCE

Belliardo, C., Koutsovoulos, G.D., Rancurel, C. et al. Improvement of eukaryotic protein predictions from soil metagenomes. Sci Data 9, 311 (2022). https://doi.org/10.1038/s41597-022-01420-4

BIO-INFORMATIQUE | BIOSTIMULATION | CHAMPIGNON MYCORHIZIEN | MÉTAGÉNOMIQUE | SOL

ARNAUD RIDEL
RÉDACTEUR
DÉPARTEMENT SPE

CONTACTS

Ces résultats sont aussi particulièrement intéressants pour la Société MYCOPHYTO dont les bio stimulants sont composés de CMA. Les CMA sont des symbiotes obligatoires que l'on retrouve naturellement en association avec plus de 90% des plantes vasculaires (Li et al., 2013). La symbiose entre plante et champignon mycorhizien améliore la croissance végétale en augmentant l'interface d'échange racinaire et, de ce fait, l'assimilation des nutriments minéraux (phosphore et azote) (Schouteden et al., 2015) ainsi que l'absorption de l'eau par la plante, en échange du carbone photosynthétique renvoyé vers le CMA (smith et al., 2010). Aussi, la mise en place de cette association réduit la sensibilité de la plante aux stress abiotiques mais aussi biotiques comme mentionné en introduction (Hijri et al., 2006; Tedersoo et al., 2014; Toju et al., 2018).

Cependant, l'utilisation des symbioses avec les CMA n'est pas à ce jour une pratique agricole de routine en raison des performances variables de cette association dépendant de l'identité des partenaires. En effet, la synergie de cette association est relative au mélange d'isolats de CMA (Crossay et al., 2019) et à l'espèce végétale impliquée dans cette relation (Eom et al., 2000) mais surtout, aux paramètres physico-chimiques de l'environnement (Campolino et al., 2022; Tedersoo et al., 2014). Mycophyto est intéressé par la compréhension des facteurs intervenant dans l'établissement de la relation hôte symbiote.

Les données métagénomiques du JGI sont accompagnées d'un ensemble de métadonnées contenant les coordonnées géographiques du lieu de prélèvement des échantillons. Sur la base de cette information, il est possible de croiser ces données avec des sources d'informations météorologiques et pédologiques pour mettre en lien la composition des communautés microbiennes et les caractéristiques environnementales du lieu de prélèvement. Ce type d'informations intéresse la société Mycophyto qui souhaite développer des outils de prédiction qui permettraient de proposer des solutions optimisées selon les caractéristiques de la parcelle agricole. En effet, la compréhension de la relation entre l'influence des

paramètres physico-chimiques de l'environnement sur la composition des communautés mycorhiziennes, et la synergie de ces communautés avec la plante hôte, permettrait d'optimiser la production et l'efficacité des produits commerciaux proposés. Plus largement, ce type d'étude peut aussi enrichir les connaissances sur l'évolution et l'adaptation des espèces microbiennes aux paramètres environnementaux.

Par ailleurs, pour extraire des informations concernant spécifiquement la distribution des CMA à partir de données métagénomiques shotgun, j'ai développé durant la thèse une méthode de metabarcoding *in silico*. Pour cela, j'ai récupéré l'ensemble des marqueurs moléculaires correspondant aux espèces du phylum Glomeromycetes dont font partie les CMA sur la ressource en ligne MAARJAM (Opik et al., 2010). La construction de profils HMM (Wheeler et Eddy, 2013) à partir de ces marqueurs a permis de scanner très rapidement les assemblages des données métagénomiques.

Parmi les différentes méthodes exploratoires développées, celle qui a fourni les meilleurs résultats concernant la représentation des CMA dans les données métagénomiques de sol est de loin la reprédiction des protéines eucaryotes. Cela confirme que cette étape est cruciale pour obtenir une bonne représentation de la diversité des communautés microbiennes.

Ainsi, les données générées durant ce travail constituent une ressource qui a été transférée à Mycophyto pour alimenter les projets de recherche et développements en cours visant à prévoir la composition optimale de mix de CMA en fonction des sols.

L'amélioration des modèles de gènes et des annotations taxonomiques ne se résume pas aux seuls CMA mais à l'ensemble des eucaryotes du sol. Ainsi, cette ressource représente un intérêt pour l'ensemble des recherches académiques visant à explorer et comprendre les communautés et fonctions des micro-organismes du sol, leurs relations et les applications qui en découlent. Cette ressource représente

également un potentiel de découverte de nouveaux gènes et nouvelles fonctions. En effet, lors de l'analyse des séquences d'ADN eucaryotes, nous avons pu identifier plus de 300 000 protéines n'ayant aucune homologie reconnaissable dans les banques de données protéiques généralistes (nr au NCBI). Bien qu'une partie de ces gènes puisse être le résultat de sur-prédictions bioinformatiques, le couplage avec des données de métatranscriptomique et/ou de métabolomique permettrait de confirmer leur existence et de lancer des investigations quant à leurs fonctions.

D'autre part, en ce qui concerne plus directement les objectifs de la thèse, ces résultats ont aussi permis d'alimenter notre librairie de référence enrichie en protéines de micro-organismes telluriques de manière plus fiable et complète pour l'étude des THG chez les NPP. En effet, nous disposons maintenant d'une plus riche représentation du pool de gènes et protéines présent dans les sols, et donc de plus de chance d'enrichir nos connaissances sur les homologues les plus proches des cas de THG chez les PPN. De plus, l'amélioration de l'annotation des gènes et protéines eucaryotes nous permettra de mieux appréhender la contribution de ces micro-organismes du sol aux événements de THG car jusqu'à présent, mis à part les cellulases de la famille GH45 d'origine fongique, il n'y avait pas d'autre exemple solide et clair de THG d'origine eucaryote chez les nématodes phytoparasites.

Chapitre II

Chapitre 2: Etude des gènes acquis par transferts horizontaux dans les génomes de nématodes parasites de plantes

1. Contexte

Comme présenté en introduction, de nombreux gènes de nématodes parasites de plantes ont été décrits comme provenant de génomes bactériens ou fongiques. La caractérisation biochimique et fonctionnelle du produit de ces gènes suggère des rôles importants dans le succès du parasitisme comme la dégradation de la paroi cellulaire des plantes ou la détoxication de xénobiotiques. Par ailleurs, des expériences d'inactivation, réalisées sur certains de ces gènes, montrent une réduction des symptômes dûs à l'infection des nématodes ce qui confirme leur importance (Béra-Maillet et al., 2000; Jaubert et al., 2002; Ledger et al., 2006; Rosso et al., 1999b). Les analyses phylogénétiques, elles, supportent l'hypothèse que ces gènes auraient été acquis par transferts horizontaux. Les gènes présumés comme acquis par transferts horizontaux sont retrouvés spécifiquement chez les espèces parasites de plantes mais rarement, voire jamais, chez d'autres animaux. Ces observations

suggèrent une importance de ces gènes dans l'adaptation évolutive au parasitisme des plantes par les nématodes.

Plus d'une vingtaine de familles de gènes différents, impliquées dans la dégradation de la paroi cellulaire des plantes mais aussi dans certains processus nutritifs, sont rapportées dans la littérature comme provenant de THG. Ces gènes étant spécifiques aux nématodes parasites de plantes mais bien souvent partagés par plusieurs de ces espèces, nous pouvons faire l'hypothèse que le gène a été acquis par une espèce de nématodes ancestrale à ces phytoparasites. Globalement, les nématodes sont retrouvés dans de nombreux lieux de vie différents. Cependant, les espèces parasites de plantes infectant les racines (i.e. la majorité des espèces connues), ont probablement évolué à partir d'un nématode ancestral tellurique. Comme mentionné aussi dans l'introduction, pour qu'un transfert de matériel génétique ait lieu entre espèces un contact direct ou indirect est nécessaire (via vecteur biologique ou moléculaire). Nous pouvons ainsi faire l'hypothèse que si des gènes ont été acquis par THG chez un ancêtre commun des nématodes parasites de plantes, ils proviennent de micro-organismes telluriques.

L'une des principales méthodes permettant d'identifier des gènes possiblement acquis par transferts horizontaux est basée sur la détection de contradiction entre l'histoire évolutive des gènes et l'arbre des espèces. Afin de détecter de telles contradictions, une recherche de similarité entre le ou les gènes du potentiel hôte et une librairie de séquences de référence représentant la diversité du vivant mais aussi les potentiels donneurs est la première étape nécessaire.

Or, les librairies de séquences généralistes utilisées jusqu'à aujourd'hui pour étudier les gènes acquis par THG chez les NPP sont très loin de représenter la diversité de gènes présents chez les micro-organismes des sols. C'est pourquoi, nous avons décidé de réaliser une nouvelles études basée sur l'utilisation d'une librairie de référence intégrant des données environnementales représentatives des

micro-organismes présents dans les microbiomes terrestres naturels. Pour cela, nous avons exploité les ressources métagénomiques shotgun disponibles publiquement afin d'enrichir notre librairie protéique utilisée pour étudier les THG.

Nous avons intégré les protéines provenant de cinq sources différentes. Premièrement, nous avons utilisé plus de 6 800 métagénomes terrestres et rhizosphériques provenant de la plateforme IMG/M du Joint Genome Institute (I.-M. A. Chen et al., 2017), l'ensemble des protéines eucaryotes re-prédites dans ces jeux de données et décrites dans le chapitre 1 (Belliardo et al., 2022), les protéines virales prédites dans ces mêmes données et disponibles sur la plateforme IMGM/VR (Paez-Espino et al., 2017), les protéines provenant de sols disponibles dans le jeu de données Mgnify proposé par l'Institut européen de bioinformatique (Mitchell et al., 2019) et enfin la librairie généraliste nr du NCBI. Par souci d'homogénéisation de la qualité des données, l'annotation taxonomique de l'ensemble des 900 millions de protéines provenant de IMG/M a également été re-déterminée en utilisant la stratégie du dernier ancêtre commun afin de garantir la fiabilité des données. La combinaison de ces ressources constitue un jeu de données de 1,4 milliard de protéines.

Après avoir constitué cette librairie de séquences représentative de la biodiversité des sols, l'objectif était donc d'utiliser cette ressource pour étudier les gènes acquis par transferts horizontaux chez les nématodes parasites de plantes. Pour avoir une vision plus complète de l'ampleur de ce phénomène, nous avons réalisé l'analyse sur 16 espèces de NPP représentant au total sept genres différents appartenant au phylum *Tylenchina*. Ces sept genres incluent majoritairement des espèces endoparasites sédentaires (sept espèces de *Meloidogyne* ou nématodes à galles et cinq espèces de la famille *Heteroderea* ou nématodes à kystes). Il est intéressant de noter que le mode de vie endoparasite sédentaire a évolué au moins deux fois indépendamment au cours de l'évolution et que les mécanismes

sous-jacents restent inconnus. Au-delà des endoparasites sédentaires, nous avons également inclus un nématode semi-endoparasite (*Rotylenchus* ou nématodes réniforme), endoparasite migrateur (*Pratylenchus* ou nématodes des lésions et famille *Radopholus* ou nématodes fouisseurs), deux nématodes des bulbes et tiges du genre *Ditylenchus* et enfin un nématode migrateur associé aux insectes et également fongivore (*Bursaphelenchus* ou nématode du pin). L'ensemble des 580 milliers de gènes prédits dans les dix-huit génomes (pour deux espèces, nous disposons de deux versions du génome) de NPP ont été comparés à notre base de données de 1,4 milliards de séquences enrichies en micro-organismes du sol afin de rechercher une indication d'une origine non animale. Cette analyse à large échelle à été permise par l'utilisation de deux logiciels automatisant des procédures de détection de possibles cas de THG et la confirmation phylogénétique des THG (Koutsovoulos et al., 2022; Rancurel et al., 2017). Ces outils ont tous les deux été paramétrés pour détecter les gènes d'origine non animale dans les génomes de NPP du clade *Tylenchina*. Après curation manuelle, nous avons détecté, au total, plus de 7 512 THG qui semblent provenir de 1 290 événements d'acquisition différents. Cela représente en moyenne une proportion d'environ 1,3% des gènes codants pour des protéines dans les génomes de nématodes parasites de plantes du genre *Tylenchoidea* et 0.5% chez *B. xylophilus*. Après détection, plusieurs analyses complémentaires ont été réalisées pour mieux comprendre l'histoire évolutive de ces gènes et leur possible impact fonctionnel sur la biologie de ces nématodes. L'ensemble de ce travail a été présenté ci-après sous forme d'article en perspective d'une publication.

2. *Article*

Mining soil metagenomes increases our knowledge of the importance of horizontal gene transfers in the evolution of plant-parasitic nematodes

Carole Belliardo[1,2], Adrien Deceneux[1], Corine Rancurel[1], Georgios Koutsovoulos[1], Marc Bailly-Bechet[1], Etienne G.J. Danchin[1]

[1] Institut Sophia Agrobiotech, Université Côte d'Azur, INRAE, CNRS, Sophia Antipolis, France
[2] MYCOPHYTO, 540 Avenue de la Plaine, 06250, Mougins, France

Abstract

Plant-parasitic nematodes (PPN) are among the most important crop pests and threaten the world's food supply. Besides the need to understand their biology to develop new control strategies, they are fascinating organisms in terms of genomic evolution. Plant parasitism has evolved multiple times independently in nematodes with some convergent evolutionary processes. For instance, all studied PPNs have acquired bacterial and fungal genes by horizontal gene transfers (HGT). Some of the acquired genes are involved in essential parasitic functions like plant cell wall degradation (PCWD) or processing nutrients from the plant. However, several major questions remain unsolved, like their total contribution to PPN gene repertoires, the functions acquired, their origin, the timing of acquisition events, and their evolutionary fate and distribution in the genomes. Most PPNs are soil-dwelling, and the underrepresentation of genes from soil microorganisms in generalist sequence libraries has previously hampered homology searches and, thus, HGT identification. To circumvent this problem, we built a protein library including over 6,800 soil metagenomes that we added to generalist protein libraries to better represent the pool of genes present in the natural environments of PPN. Then, we performed an HGT detection on protein sets from 18 plant-parasitic nematode genomes of the *Tylenchina* clade, constituting the most economically important and highly diverse group of PPN, against our library enriched with soil proteins.

After manually curating automated HGT detection, we finally identified 7,512 HGT candidates related to 1,208 acquisition events, including 514 events common to several genomes. The proportion of genes acquired by horizontal transfers with phylogenetic confirmation in PPN genomes ranges from 1.38% of protein-coding genes in *Tylenchoidea* and 0.54 in *Aphelenchoidae* PPN families. HGT mainly originates from bacteria, but we also observe fungi, protists and for the first time plants. The usage of metagenomic data clarified the history of previously described HGTs but also identified hundreds of new HGTs. Functional analyses of the newly identified HGTs indicate a wide diversity of potential functions whose biological implications can be more precisely described in in-vitro experiments. Integrating environmental data in our reference library has allowed us to extend the detection of HGTs and to complete the catalogue of potential donor offspring.

Keywords Plant-parasitism | Nematodes | Horizontal gene transfer | Metagenomics | genome evolution | parasitism.

1. Introduction

To address ecological challenges in the context of population growth, we must develop more environmentally-friendly yet efficient crop protection methods while increasing food production. Plant-parasitic nematodes (PPN) are globally distributed and among the most important crop pests threaten the world's food supply. They infect most vascular plants, including an extensive range of species with agronomic interest, and are responsible for 11% of crop losses yearly (Abad et al., 2008; Singh et al., 2015). Therefore, it is crucial to understand how these species have evolved the ability to parasitise plants to propose more evolutionary-informed and efficient methods to control them and reduce the agricultural losses they are responsible for.

Currently, more than 4,000 species of nematodes have been described as plant parasites (Decraemer, Wilfrida and Hunt, David J, 2006; Maule and Marks, 2006). These species are currently distributed across four orders of the phylum Nematoda: *Triplonchida* (clade 1), *Dorylaimida* (clade 2), *Aphelenchida* (clade 10 b), *and Tylenchida* (clade 12) in the phylogeny proposed by van Megen (van Megen et al., 2009). In all phylogenetic analyses of Nematodes, plant parasitism appears to have emerged independently multiple times (Ahmed et al., 2022; Holterman et al., 2017; van Megen et al., 2009). Due to their colossal food damage and their high economic impact, *Chromadorea* class comprises the most extensively studied nematode species, including *Tylenchida* (clade 12) and Aphelenchoida (clade 10 b), which together represent most of the living and feeding patterns of plant-parasitic nematodes (Danchin et al., 2017). All phytophagous species face similar challenges from the plant defences and feeding processes; thus, the evolution towards a plant-parasitic lifestyle requires specific morphological and biochemical characteristics (Holterman et al., 2017). To circumvent plant defences, PPN taxa seem to have followed convergent evolutionary trajectories (Bird et al., 2015), involving both vertical inheritances coupled with molecular evolutionary processes and horizontal transfer of genes (HGT).

Once deemed a quirk, horizontal gene transfers in eukaryotes, including animals, are now being more and more described in the literature (Danchin, 2016). The evolution of species requires genetic variations of individual features, which may come from gene flow, transmitted mainly by vertical inheritance. However, variations in the DNA material are not only due to the vertical inheritance of mutations or even the birth of *de novo* genes but also result from the incorporation of foreign genome chunks through lateral transfer that the offspring could vertically inherit. We know these events play an important role in the evolution of prokaryotes and unicellular eukaryotes. Gene flows between microbial cells provide new features for fighting biotic or abiotic attacks, developing new skills for adapting to the environment, or access to new ecological niches (Alačević, 1963; Arnold et al., 2022). Recent studies have shown that transferring genetic material between species has also occurred in genomes of multicellular organisms, including animals (Boto, 2014; Leger et al., 2018). This process is relatively uncommon for metazoan species because to be integrated and transmitted, the acquired genetic material must reach the nuclear DNA of germ cells which are compartmentalised from the rest of the organism. Some detailed mechanisms have been proposed, including the delivery of DNA chunks into the cell (e.g., viruses or lipid vesicles), entry into the nucleus (e.g., molecular interaction with nuclear proteins), and the host DNA insertion (e.g., transposable element activity or DNA breakdown). Viruses are excellent vessels for DNA which can be transported between individuals and/or invade host genomes. Some viruses are experts in introducing DNA into genomes; this mechanism is very well described for retroviruses with an active mechanism to insert their genetic material into the host genome through an enzymatic arsenal (Whitcomb, 1992). However, the viruses belonging to this

family are not the only ones that can integrate into the host cell genome, as several studies show endogenous viral elements in host genomes (Gilbert and Belliardo, 2022; Pienaar et al., 2022). Finally, because their replication occurs within the host cell, they can capture genetic material from the host. Currently, the transport of genetic material between animals through viruses has been demonstrated experimentally in Drosophila (Loiseau et al., 2021).

DNA transfer mechanisms between microbes and animals have yet to be wholly elucidated or experimentally assessed. However, close contact between donor and recipient organisms, such as symbiotic or trophic interactions, could be a facilitating element. Similarly, infection with a virus or parasite could just as efficiently serve as a vector for transfer between species, even if it is not an intimate relationship.

Although these biological events are not fully explained and are relatively scarce, a growing body of research shows that they appear to have significantly impacted the evolutionary history of animals through the emergence of new abilities (Husnik and McCutcheon, 2018), such as plant parasitism function (Danchin et al., 2010). Strikingly, some HGT discovered in plant-parasitic nematode genomes involves biochemical processes essential for parasitic functions. Indeed, one of the first interkingdom HGT cases described from bacteria toward metazoan involved PPN species and genes encoding glycoside hydrolase enzymes with cellulase activity in the cyst nematode genome (Smant et al., 1998). These enzymes give nematodes the ability to degrade the main components of the plant cell wall. Historically, endogenous glycoside hydrolase activity in animals was reported for the first time early in 1960 in nematodes (Dropkin, 1963). Only after several decades genes encoding enzymes with a cellulase activity were described in two cyst nematode species, and the bacterial origin of these genes seemed to be the most likely hypothesis, given the sequence similarity between these genes and those found in soil bacteria (Smant et al., 1998, Keen and Roberts, 1998). Enzymes with the same biological function coming from the same gene family were then discovered in root-knot nematodes (Rosso et al., 1999). Then, diverse cases of putative HGT were described in PPN, including genes involved in the modulation of the plant's defence systems, the establishment of a nematode feeding site, and the synthesis or processing of nutrients (Craig et al., 2009; Danchin, 2016; Danchin et al., 2010; McCarter et al., 2003; Opperman et al., 2008). Interestingly, similar evolutionary trajectories are observed in other plant parasites, such as glycosides hydrolase enzymes acquisition from fungal donors in phytophagous beetles (Kirsch et al., 2014) or even HGT involved in the detoxification function in the plant leaves feeder whitefly that comes from host plant genomes (Lapadula et al., 2020). Altogether, these observations support the hypothesis of a strong impact of HGTs in the evolutionary history of some animals that became notorious plant pests. Today, all PPN species analysed at the omic level have been shown to contain genes from fungal or bacterial donors (Craig et al., 2009; Danchin et al., 2017, 2010, 2010; Eves-van den Akker et al., 2016). Moreover, transcriptomic, biochemical and proteomic data support the functional 'domestication' of genes derived from HGTs, which seem to have played an essential role in the multiple independent emergences of plant parasitism in nematodes (Danchin et al., 2010; Jaubert et al., 2002). These horizontally acquired genes seem to come from several independent transfer events because even in the case of genes coding for the same protein function, the genes seem to originate from different donors according to the nematode clade (Noon and Baum, 2016). For example, in Tylenchida genomes (clade 12), enzymes with cellulase activity belonging to the glycoside hydrolase family 5 (GH5) are commonly found in bacterial genomes, but cellulases found in the genomes of *Aphelenchida* (clade10 b) species belong to a completely different family (GH45) and seem to have been acquired from fungi by horizontal gene transfer

(Haegeman et al., 2011). Thus, it becomes clear that multiple horizontal gene acquisition events have occurred in PPN genomes. Since many of them are related to parasitic functions, such events have probably played a crucial role in the adaptive evolution of nematodes towards a plant-parasitic lifestyle. Moreover, the so far systematic detection of HGT in PPN lineages suggests that HGT event is a prerequisite for plant-parasitism success. Nevertheless, several questions remain unsolved: What is the total contribution of HGT to gene repertoires in PPN genomes? What kind of functions have been acquired via HGT? Which are the donor organisms? Furthermore, when did these evolutionary events occur in the evolutionary history of the Nematoda phylum?

Approaches to detect horizontal gene transfers between distantly related organisms classically start with identifying genes returning higher sequence similarity with distantly related clades than phylogenetically closer clades (Koutsovoulos et al., 2022; Rancurel et al., 2017). To discover this peculiarity, sequences of the organism of interest need to be compared to an as comprehensive as possible reference sequence library containing both sequences from putative donors and more closely related species. One major limitation in previous HGT studies is the underrepresentation of soil-dwelling microorganisms (i.e. potential donors) in generalist sequence libraries (Parks et al., 2018, 2017; Watson, 2018) despite these species being possible donors. In recent years, the rise of metagenomics has allowed the discovery of previously unknown microorganisms and expanded our understanding of the genetic diversity of microorganisms in many different complex environments, including soil and plant-associated microbiomes (Fierer, 2017; Ramirez et al., 2014; Taberlet et al., 2012). Recent efforts have focused on the *de novo* assembly of bulk metagenomic sequencing reads into metagenome-assembled genomes (MAGs) or contigs, uncovering the genetic content and informing on the molecular functions of uncultured microorganisms (Nayfach et al., 2021; Naylor et al., 2020). These advances in microbiology have highlighted the gap between the high genetic diversity of microorganisms in natural environments and the poor representation in the standard sequence libraries and paved the way for a more efficient and precise detection of HGT in PPN. In addition to limitations on the explored protein dataset, previous attempts to characterise HGT at a whole-genome level in PPN were either restricted to only a few organisms from the same genus (Paganini et al. 2012) or were based on homology search results alone without phylogenetic validation and thus prone to substantial rates of false positives (Lai et al., 2022).

To circumvent these limitations, we assembled an as comprehensive as possible reference library with metagenomic data from four public resources (Belliardo et al., 2022; Chen et al., 2017; Mitchell et al., 2019; Paez-Espino et al., 2017), then combined them with the generalist protein library NCBI *nr* (O'Leary et al., 2016). The first challenge was manually curating and integrating these massive datasets into a usable protein library. Then, we searched for candidate HGT on proteins predicted from 18 plant-parasitic nematode genomes of the *Tylenchina* clade, covering seven different genera and including both migratory and sedentary parasites. Thus we compared a highly diverse group of PPN against our library enriched with soil proteins. We then confirmed the horizontal acquisition of putative HGT using automated phylogenetic analyses (Koutsovoulos et al., 2022).

2. Materials and Methods

2.1. Data collection

2.1.1 Plant-parasitic nematode proteins

We downloaded predicted proteomes from 18 plant-parasitic nematode genomes (Grynberg et al., 2020). The number of proteins per dataset ranged from 10,895 (*Meloidogyne graminicola*) to 101,269 proteins (*Meloidogyne arenaria*), yielding 598,842 proteins. We studied 17 species from the superfamily Tylenchoidea and one from Aphelenchoidea (Table 1, Supplementary data table 1).

Table 1 | List of plant-parasitic nematodes genus studied

nb.	Common name	Scientific genus	Lifestyle
			17 Tylenchoidea superfamily
7	Root-knot nematode (RKN)	*Meloidogyne*	obligatory sedentary endoparasite
5	Cyst nematode (CN)	*Heteroderidae**	obligatory sedentary endoparasite
2	Stem and bulbs nematode (SBN)	*Ditylenchus*	obligatory migratory endoparasite
1	Lesion nematode (LN)	*Pratylenchus*	obligatory migratory endoparasite
1	Reniform nematode (RN)	*Rotylenchus*	obligatory migratory semi-endoparasite
1	Burrowing nematode (BN)	*Radopholus*	obligatory migratory endoparasites
			1 Aphelenchoidea
1	Pine wilt disease nematode (PWN)	*Bursaphelenchus*	facultative migratory

**(Heteroderidae family - including three from the Globodera genus and 2 of the Heterodera genus).*

2.1.2 Reference protein library enriched with soil metagenomic data

First, we downloaded publicly available assembled soil metagenomic data from shotgun sequencing reads of the two leading platforms specialised in analysed microbiome sequencing projects, IMG/M and Mgnify, respectively hosted by the Joint Genome Institute and the European Molecular Biology Laboratory (Chen et al., 2017; Mitchell et al., 2019). The European server provided a single file of pooled proteins and their taxonomic annotation resulting from a last common ancestor assignment approach. Hence, we filtered proteins from soil samples or associated them with plant roots using metadata. In contrast, the most extensive resource, the JGI server, provides metagenome datasets containing downloadable raw and pre-processed files by sample. Thus, we collected metagenomes of 5,988 'Terrestrial' samples in the environmental metagenomes category and 884 plant-associated metagenomes in the host-associated category (available data 2020, October; Supplementary Data (Belliardo et al., 2022)). The data acquisition was performed via the IMG/MER Cart genome portal. The joint genome institute also provides the IMG/VR database, a complementary and specific annotation of viral genomes identified from metagenomes (Paez-Espino et al., 2017), that we appended to our reference library.

In the same way that conventional pipelines can miss the annotation of viral sequences in environmental data because they are focused on prokaryotes, sequences from eukaryotic organisms can also be wrongly annotated despite their importance in terms of biomass and ecological functions.

To better represent the natural biodiversity of soil microbiomes, including eukaryotes, we added proteins *de novo* predicted and re-annotated in the same 6,872 soil metagenomes from IMG/M of the JGI with a eukaryotic tailored workflow (Belliardo & al, 2022). Finally, we completed our protein library with the January 2020 release of the NCBI nr database. The sum of all resources yielded more than 11 billion proteins corresponding to almost 2 TB of data.

2.2 Integration of protein sources into a single library: from metagenomic data to a reliable protein database

2.2.1 Data curation and quality control

Proteins from metagenomes were collected from different servers and originated from automated prediction and annotation pipelines resulting in highly heterogeneous quality. Therefore, the challenge was to make this massive data more reliable and meaningful. The most extensive metagenomic datasets originated from IMG/M, providing predicted proteins and genomic data (IMGM: 1,4To / Mgnify: 272Go, Table 2). The mining of IMG/M MAGs revealed high heterogeneity of the contig lengths between metagenomes due to variations in sequencing technologies, experimental protocols, pipeline version used and biological features. Because most data initially consisted of short sequencing reads and a low assembly success rate, half of the contigs were shorter than 296bp (Belliardo et al., 2022). These short contigs increase the volume of data to process and are unlikely to contain complete genes, hence providing no more information on gene diversity (Brent, 2007). We first applied stringent filters on IMG/M soil metagenomes based on the assembly quality by keeping only proteins from contigs at least 1Kb long or containing at least three genes predicted by Prodigal in JGI files. Moreover, a minimal length threshold of 50 amino acids for all metagenomic datasets (Nevers & al., 2021) was also applied. Filters reduced by almost a factor of 4 the volume of the dataset and further increased the accuracy of the data quality.

2.2.2 Reducing protein redundancy

Although filtering significantly reduced the amount of data, we expected a remaining redundancy in the pool of metagenomic proteins due to data originating from thousands of independent studies. Furthermore, some sequencing data came from time points sampling the exact location. Redundancy at the protein sequence level had to be eliminated to obtain a quality dataset similar to that of the standard non-redundant libraries.

We pooled proteins by metagenomic source (IMG/M, MGnify, IMG/Vr) and clustered each set with the Linclust software of the MMseq2 metagenomic toolkit (Steinegger et al., 2018). We used the "*easyclust*" workflow to perform the clustering with at least 99% sequence identity and covering at least 90% of the target. Clustering was performed individually on libraries, but redundancy between different libraries was conserved. Overall, the number of sequences was reduced by 30% (Table 2).

2.2.3 Enhancing the reliability of taxonomic information

In the most extensive metagenomic public repositories IMG/M, the taxonomic assignment is solely based on the best BLAST hit against NCBI *non-redundant* (*nr*) regardless of the percent identity (Chen et al., 2017) resulting in unreliable information. To obtain a more reliable taxonomic annotation, our strategy was to apply the last common ancestor algorithm of a Diamond search against the NCBI's nr using the strategy described in (Belliardo et al., 2022) to the 924 million soil metagenomic proteins from the JGI. In the LCA mode, Diamond will assign an NCBI taxonomic identifier (i.e. TaxID) based on the last common ancestor of all the hits with a score not diverging by more than 10% from the best hit score (Buchfink et al., 2015). Using an LCA approach constitutes a substantial gain in taxonomic annotation reliability compared to approaches based on the best BLAST hit alone, this single best hit being potentially miss-annotated itself or sharing only a common identity with the query sequence.

We combined the almost billions of curated and taxonomically improved metagenomic proteins with the NCBI nr databases to cover a wide spectrum of living organisms. Then, we built a Diamond formatted protein database with default parameters. Finally, we obtained an improved and non-redundant comprehensive protein

database more representative of the natural soil biodiversity.

2.3 Detection of HGT in PPN, Genome screening: identification of putative HGT

2.3.1 Identification of possible HGT based on homology

To detect putative HGT, we used Alienness (Rancurel et al., 2017) and AvP (Koutsovoulos et al., 2022), which automatically computes HGT metrics based on sequence similarity search results (i.e. BLAST or DIAMOND). This software was tuned to exclude hits with *'Tylenchina'* proteins (TaxID: 6300) and keep as a taxon of interest *'Metazoa'* hits (TaxID: 33208). With this configuration, those softwares will detect putative HGT of non-metazoan origin in *Tylenchina*.

The first metric calculated by Alienness is the Alien Index (AI) based on the difference between e-values of the best donor hit and best recipient hits, which was first described in (Gladyshev et al., 2008). Since e-values are sensitive to the database size, the AI scores from different libraries cannot be compared. However, because they are based only on the best Metazoan / non-Metazoan BLAST hits, they are highly sensitive to database contamination or taxonomic misannotation. If the subject protein of the best hit is erroneously assigned a non-metazoan taxID in the database, this metric will wrongly detect a candidate HGT. In contrast, erroneously labelling with metazoan taxID to a non-metazoan protein of the database will not enable HGT detection. The second metric calculated by AvP (Koutsovoulos et al., 2022) was proposed to circumvent this issue by aggregating all normalised bit-scores of donor and recipient sequences before calculation. This Aggregate hits score (AHS) will decrease misdetection related to taxonomic annotation errors in databases. To collect as many putative HGT cases as possible at this step, we selected all PPN proteins returning either a positive AI or AHS (i.e. having higher similarity with non-Metazoan than Metazoan proteins). Briefly, similarity analyses were conducted independently for each PPN species with the software DIAMOND, designed for high-performance analysis of protein sequences across libraries.

All predicted proteins are compared against our comprehensive protein library using DIAMOND in BLASTp mode (Buchfink et al., 2015) with a *max-target-seqs* number of 500. Then, we used Alienness and AvP to parse DIAMOND results using a homemade taxonomic cross-reference combining NCBI's taxonomy and our metagenomic taxonomy assigned as previously described.

A penalty e-value or bit score (1 for AI and 0 for AHS) was automatically assigned to the best metazoan or non-metazoan score when no BLAST results were found in these taxonomic groups. To allow detection of HGT that took place in ancestors of PPN, self-BLAST results to *Tylenchina* were ignored for the calculation of AI and AHS. Moreover, all hits with proteins assigned to an 'unclassified' (TaxID: 12908) or 'Other' (TaxID: 28384) taxonomy were also excluded from the calculation. Thus, no index values could be calculated for proteins returning no significant hit or only hits to 'unclassified 'and 'Other' TaxID. In contrast, for both AI and AHS a positive value (> 0) reveals a better match with non-metazoan species than with metazoan species, which might indicate possible HGT (Koutsovoulos et al., 2022; Rancurel et al., 2017).

2.3.2 Clustering of orthologous proteins and phylogenetic validation

When a PPN protein is more similar to non-metazoan proteins, this indicates a gene possibly originating from horizontal transfer. We performed phylogenetic analysis for all HGT candidates to confirm their origins and provide a complete evolutionary history.

For the following steps, we used the AvP toolkit (Koutsovoulos et al. 2022), which provides a suite of software which automatically performs all common steps of phylogenetic HGT analysis and validation. In summary, AvP allows retrieving the homologous sequences, producing multiple

alignments, constructing the phylogeny and analysing the topology to assess whether it supports the HGT hypothesis.

2.3.2.1 Clustering of homologous proteins

Genes acquired by a common ancestor of PPNs are expected to be found in descendant species resulting in orthologous relations between genes. Furthermore, after their integration, HGT can be duplicated into the host genome resulting in paralogous genes. Thus, to recover this information, the first step after the detection of putative HGT is to cluster / group homologous PPN proteins with their respective hits in the reference databases.

To cluster homologous proteins, we used a homemade python script (see below) performing a meta-clustering by integrating the results of 3 approaches of homology analyses:

- Single Linkage clustering of HGT performed with **AvP** *'prepare'* module run in default mode(Koutsovoulos et al., 2022). This module groups the query species sequences based on the percentage of shared blast hits (by default, 70%).
- Phylogenetic orthology inference performed with **OrthoFinder** (Emms and Kelly, 2015)**.** This software determines the correspondence between genes in different organisms. We used publicly available Orthofinder results from (Grynberg et al., 2020), https://doi.org/10.15454/IIAQOW.
- Clustering by shared protein family domains. For this, protein family domain annotation was performed with **InterProScan** (Jones et al., 2014)**,** and we selected only Pfam annotations.

First, we aggregated the groups sharing an AvP and/or OrthoFinder group into supergroups. Then, we also aggregated supergroups with the same most represented PFAM domain. We retained only the largest domain for each protein for this last step.

2.3.2.2 Phylogenetic analysis

Phylogenetic analysis was automatically performed with AvP software using our comprehensive protein library as input. This analysis was performed by re-using the AvP *'prepare'* module, providing a grouping file from our meta-clustering of homologous proteins. This procedure automatically collects all protein sequences of groups of queries and significant hits from the database. First, AvP created one Fasta file by orthologous supergroups. Then, we ran the AvP 'detect' module with the trimming option disabled and minimal *node_support*' value of 75, under which branches were collapsed into polytomies to only retain highly-supported branches. AvP parallelise the phylogenetic reconstruction by homologous supergroups by aligning each multi-protein fasta file using MAFFT (Katoh and Standley, 2013), and then reconstructing phylogeny using FASTTREE (Price et al., 2010). After phylogenetic reconstruction, AvP automatically classifies queries according to the tree topology based on the selected taxon of interest (TOI, here Metazoan: taxID 33208) and excluded group (EG, here *Tylenchina*: taxID 6300). To deal with the potential erroneous taxonomic assignment in the reference library, we configured AvP to accept a maximum of 20% of metazoan proteins in sister branches. Beyond this threshold, the putative HGT is labelled 'complex' to remain stringent.

2.3.2.3 Manual curation of miss HGT classifications

Some pitfalls could remain in automated HGT analysis related to complex cases, such as (i) host genome and (ii) reference library contaminations or (iii) independent HGT events that happened in the ancestor of relatively close taxa. A HGT event here defined has a monophyletic PPN clade surrounded by more than three non metazoan proteins. Thus, we considered four additional features described below to validate HGT based on phylogenetic topologies:

- The percentage identity between HGT candidates and their best non-metazoan hits reported by Alienness (Rancurel, 2017)
- The orthology conservation between several PPN genomes based on detecting protein from other *Tylenchina*

genomes in branches related to a unique transfer event.

- The content of sister branches requires at least three non-metazoan proteins in sister branches filtered by a custom python script (see below)
- The genomic environment, which is calculated by the 'local_score' module of AvP and configured to compute the score on ten surrounding upstream and downstream HGTs. We labelled as contamination any species-specific HGT without any surrounding host gene.

The first feature revealing potential contaminations or erroneous taxonomic annotation in protein libraries is the high percentage identity with homologous genes from possible donors, as mentioned in (Ku and Martin, 2016), where authors suggested a threshold of 70 percent identity to discriminate HGT from contamination. Also, putative HGT classification for proteins predicted on DNA chunks composed of a high density of genes matching only with distant (donor) species suggests contamination (Koutsovoulos et al., 2016). We analyse the gene environment for all PPN species except *Rotylenchulus reniformis* due to missing GFF files. Genes with more than 70% identity with putative donors or without nematode genes in the range of 10 genes before or after and no ortholog in any other PPN genomes are considered as resulting from contamination. In addition, putative HGT with less than three non-metazoan species in sister branches was considered indicative of erroneous taxonomic classification in the reference library.

3. Identification of putative donors

All metagenomic proteins were taxonomically annotated using the Last Common Ancestor function implemented in Diamond. Those from IMG/M and IMGMeuk were annotated with default parameters and the January 2020 release of the NCBI nr database as a reference library. For data from Mgnify and IMGVR, we used the original annotations provided. Before pooling, proteins were labelled according to their database source '_MG_IMGM', '_MG_IMGMeuk', '_MG_IMGVR':, '_MG_MGY'].

This analysis used trees reconstructed by AvP in newik format. Information about putative donors was extracted from sister branches of the HGT event (custom script, see below). For each HGT, we browse the tree from the leaf to the root, searching for a node spanning a monophyletic set of leaves that we call an "HGT event" which includes all homologous HGT followed by robust branches defined as containing at least three non-metazoan proteins (Fig. 1). Among HGT event branches, non-*Tylenchina* leaves isolated among a set of nematode proteins are considered as a mis-annotation in the reference library and ignored. In contrast, successive branches containing more than three non-metazoan proteins are deemed to be robust branches. Then, for each protein of sister branches, we assigned a taxonomic kingdom annotation, including 'Bacteria', 'Archaea', 'Viruses' and 'Eukaryota', declined into 'Eukaryota@Metazoa' 'Eukaryota@Fungi', 'Eukaryota@Sar', 'Eukaryota@Viridiplantae' and 'Eukaryota@other_non_metazoa'. If more than 80% of branch kingdoms are similar we attribute a branch label and a LCA of those lineages are performed to refine donor identity. Otherwise, the "NM_mixture " name is attributed.

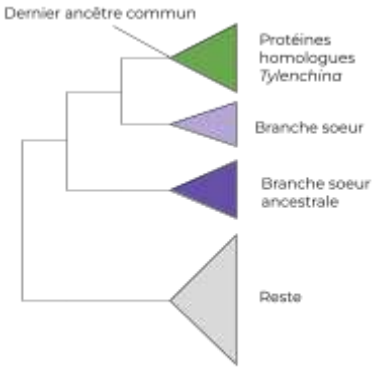


Schéma de la topologie des arbres phylogénétiques

4. Functional annotation and Gene ontology terms enrichment

To identify conserved protein domains, we annotated, for the 18 studied *Tylenchina* genomes, all predicted proteins with InterProScan (Blum et al., 2021; Jones et al., 2014), with the option -iprlookup -gotermsemployed to assign Gene Ontology (GO) terms from the identified InterPro domains. Then, we determined whether some GO terms were significantly enriched in HGT candidates compared to the background of the species analysed using the hypergeometric test implemented in GoFuncR software v.1.10.0 (Grote, 2017) using default version 2020, March 23 of the GO-graph database.

We used the 'refine' option to retain non-redundant general terms (higher in the hierarchy, usually inferred by FUNC) when most of the corresponding proteins in a node were spread out across daughter GO terms. When overlaps between proteins associated with several enriched GO were identified, these proteins were joined into representative groups. Each group was given a more general term that represents and/or explains the biological processes and/or molecular functions in which they participate. GO terms fold-enrichment (FE) values were calculated as the ratio between the observations and the expected frequency of genes.

3. Results

3.1 Comprehensive and accurate soil metagenomic protein library

We hypothesised that genes acquired by horizontal transfers in plant-parasitic nematode genomes mostly originated from soil-dwelling microorganisms. In such a scenario, knowledge of uncultured soil microorganisms could increase our understanding of the impact of HGT on the gene repertoires and evolutionary history of phytoparasitic nematodes. To figure it out, we populated our reference library with proteins predicted in terrestrial and rhizospheric metagenomes.

We downloaded proteins from 6,872 public metagenome-assembled genomes (MAG) studies publicly available on the IMG/M platform and the pooled set of proteins from the Mgnify platform. The first step was making this massive data more reliable for protein prediction and subsequent taxonomic assignment to ensure robustness in HGT analysis.

We filtered the data according to assembly quality and sequence completeness at the contig and protein levels. Thus, we only retained high-quality protein predictions on contigs from terrestrial and plant-associated metagenomes. This quality filtering reduced metagenomic datasets by tenfold (85% for the IMG/M and 99% Mgnify), removing 9.6 billion proteins (Table 2). To further optimise the time needed to mine this library, we have also reduced the redundancy by clustering similar proteins, yielding 950 million non-redundant proteins (Table 2). Overall, we have gained accuracy and compactness without information loss and obtained a high-quality and non-redundant dataset of the 950 million proteins from soil-dwelling and rhizospheric organisms complementing the 462 million proteins from NCBI *nr* (Table 2).

The taxonomic classification of proteins from uncultured microorganisms is challenging and could constitute a pitfall in detecting HGTs. Indeed, the taxonomic assignment in JGI automatic workflows is based on the single best BLAST hit against different versions of the NCBI's nr depending on the release date of the metagenomic data (Chen et al., 2017; Mitchell et al., 2019). This automatic taxonomic annotation is highly sensitive to taxonomic annotation errors because it only relies on the best blast hit without a similarity threshold. If the taxonomic annotation of this hit is erroneous due to contamination or manual error, it is automatically propagated without further quality check.

A more reliable method for taxonomic assignment of sequences distantly related to those of known organisms, such as ancient or current metagenomic data, is usually based on a last common ancestor approach, taking into account multiple hits. This method was

previously used for the taxonomic annotation of soil eukaryotic proteins (Belliardo et al., 2022) and for Mgnify dataset annotation (Mitchell et al., 2019).

We redid the taxonomic assignment for the entire metagenomic protein library with a last common ancestor approach and up-to-date reference NCBI nr library to homogenise the quality and reliability between the different data sources (i.e. IMG/M vs Mgnify). This process resulted in a taxonomic annotation with variable ranks (deeper or shallower) between proteins depending on the consistency of the n best hits taxonomic annotation with less than 10% score drop with the best hit. This assignment method contrasts with the original JGI annotation, which always gives information at the deepest rank, i.e. species or strain, except for unclassified or other sequences, regardless of the score and only based on the single best hit.

Integrating proteins from five different sequence resources from IMG/M (Chen et al., 2017), IMG/VR (Paez-Espino et al., 2017), IMG/euk (Belliardo et al., 2022), Mgnify (Mitchell et al., 2019) and NCBI nr (Benson et al., 2012), our custom library contained a total of 1.4 billion proteins and constitutes a resource three times larger than libraries used in previous HGT analyses (Table 2). In addition, metagenomic data significantly increased the taxonomic diversity of the protein library.

Indeed, 82% of protein taxa are assigned to ranks from genus to kingdom. This taxonomic assignment to more ancestral ranks reflects the absence of multiple hits from closely related species in the NCBI's nr and highlights significant enrichments of the soil biodiversity representation. Thus, populating a reference library with all available soil metagenomic data provided an unparalleled resource for HGT study in soil-dwelling organism genomes.

Table 2 | Reference protein library information according to data collection

Protein library	Nb. of protein	size (Go)	Filtering	size (Go)	Clustering	Size (Go)	Ref.
Mgnify	1,106,951,200	272	12,783,475	2.3	11,608,667	1.7	Mitchell et al., 2019
IMG/M	9,919,643,025	1400	1,520,218,604	338	924,675,615	229	Chen, 2017
IMG/VR	68,018,527	19	4,769,719	1.3	2,945,078	0.8	Paez-Espino et al., 2017
IMG/euk	8,000,000	3	6,500,000	2.6	4,771,679	1.9	Belliardo et al., 2022
NCBI *nr*	462,834,406	223	462,834,406	223	462,834,406	223	(Benson et al., 2012)
Total	**11,565,447,158**	**1,917**	**2,007,106,204**	**567**	**1,406,835,445**	**446,4**	N/A

3.2 HGT content in Tylenchina genomes using environmental data

Using our soil protein library, we performed HGT detection and analysis on the full proteome of a wide range of PPN species to understand the impact of HGT in their evolutionary history. Thus, we screened proteins from 18 genomes of *Tylenchina* species including seven root-knot nematodes (i.e. *Meloidogyne* genus), five cyst nematodes (i.e. *Heteroderidae* family), two stem and bulbs nematode (i.e. *Ditylenchus* genus), one lesion nematode (i.e. *Pratylenchus* genus), one reniform nematode (i.e. *Rotylenchus* genus), one burrowing nematode (i.e. *Radopholus* genus) and one pine wilt disease nematode (i.e. *Bursaphelenchus* genus) (Table 1).

HGT events have already been described in the genomes of these species through independent studies, generally focused on a single gene and not always including phylogenetic validation. Here, we applied a reliable phylogenetically-based HGT detection and analysis method to all PPN species to provide a more robust and comprehensive history of HGT in phytophagous nematodes of *Tylenchina* and *Bursaphelenchus* genus.

3.2.1 Initial detection of putative HGT using similarity metrics

Horizontal gene transfer arises from the transmission of genetic material between a donor and a recipient organism that may be evolutionarily distant, even across kingdoms. We were interested in HGT of non-metazoan origin in plant-parasitic nematodes in this study.

To perform this analysis, we used Alienness software (Rancurel et al., 2017) on the 18 PPN genomes. This is a fast, high-throughput initial approach to identify a gene as a potential HGT, based on a similarity metric between the most similar genes in closely and distantly related species. Screening the 598,842 proteins from 18 *Thylenchina* genomes against our custom library revealed 43,042 proteins resembling more non-metazoan than metazoan proteins. Of these putative HGTs, 16,560 cases constitute high-presumptive HGT cases by displaying positive values for both AI and AHS metrics.

Putative HGTs are distributed between the 18 genomes and represent, on average, 6.95% of predicted proteins per genome, with a minimum value of 3.31% (727 putative HGT) observed for *P. penetrans* and a maximum of 10.10% (6,036 putative HGT) in *M. incognita Newton* (Fig 1, supplementary data table 2).

3.2.2 Phylogenetic confirmation and analysis of putative HGT

Similarity-based metrics help detect putative HGT. However, this approach does not always inform on the evolutionary history of genes. To confirm putative HGT cases, we need to reconstruct the relationship between homologous genes by phylogenetic analyses, including a broad range of homologous sequences.

To validate the origin of putative HGT in PPN genomes, we used AvP software (Koutsovoulos et al., 2022) which automatically computes phylogenies and analyses tree topologies. AvP was configured to perform phylogenetic reconstruction for genes with more than three hits, yielding 42,525 putative HGT cases that were further phylogenetically analysed. The first step to managing a phylogenetic analysis at this multi-species scale is to define groups of proteins deemed homologous to reconstruct comprehensive trees and more readily assess the number of horizontal acquisition events. To retrieve homologous HGT cases, we developed a meta-clustering method based on three similarity signals, associating single linkage clustering based on a majority of shared BLAST hits (i.e. 11,979 groups), orthology relationship deduced from MCL clustering of reciprocal best-hits (i.e. 4,788 groups) and shared protein families domains (i.e. 1,882 groups). Using as a gold standard a manual expert CaZyme annotation of the *M. incognita* proteome performed by the glycogenomics team, we compared different method clustering in succeeding to aggregate genes from the same gene family.

The meta-clustering method gives more consistent results to cluster putative HGT from the same gene family into a single group, relative to any of the three methods used alone. For example, HGT encoding polysaccharide lyases proteins with pectate lyase activity of the PL3 family previously described in (Haegeman et al., 2011) were clustered into 14 different groups by the linkage clustering approach and into height different groups by the orthology analysis. However, our meta-clustering method combined them into one more extensive group.

Using our meta-clustering methods allowed the grouping of the 42,525 putative HGT from the 18 species into 2,762 orthologous groups, with sizes ranging from one to 6,547 HGT per group, with an average of 15 HGT genes. Then, AvP automatically reconstructed phylogenies for each group and classified putative HGT according to tree topologies.

Setting the software to be tolerant to some miss-annotation in the reference library (methods), the hypothesis of the horizontal acquisition was confirmed with strong support from topologies for 7,565 protein-coding genes spread in 950 trees, which encompassed 1,180 acquisition events. Therefore, several trees include multiple independent HGT events. A given HGT event can include multiple PPN proteins from multiple species (potential acquisition in an ancestor) or the same PPN species (duplications after acquisition).

The software could not conclude for 9,411 putative HGT classified as 'COMPLEX' due to more than 20% of metazoan proteins in a sister or ancestral sister branch. Considering potential independent acquisitions in other metazoan taxa, these genes related to 1,272 events in 532 trees were labelled 'complex' in this first tree topology analysis.

3.2.3 Ruling out false positive detection to refine the HGT identification

The automatic processing with AvP software coarsened the identification of HGT by rapidly selecting robust cases. However, automated approaches cannot wholly resolve complex cases such as contamination in (i) host genome or (ii) reference library. These biases related to the nature of data could not be automatically managed by HGT software and required more expertise. One-third of the complex cases were randomly selected and manually analysed to determine whether general rules could refine phylogenetic validation of HGT in PPN nematode genomes using a custom library populated with metagenomic data.

The first feature revealing possible contaminations in either host or library proteins is the high percent identity between candidate HGT and homologous genes from possible donors, as mentioned in (Ku and Martin, 2016). Another gene feature potentially indicative of the presence of contamination is the genomic environment of the putative HGT. Indeed, an HGT case related to a gene located on a DNA chunk composed entirely of genes classified as HGT candidates will be strongly presumed as contamination.

In contrast, the detection of endogenous genes from the host genome in the genomic environment of HGT or the occurrence of a homologous HGT in several genomes makes the contamination hypothesis less likely. These last features support the hypothesis of horizontally acquired genes.

Applying stringent rules based on these assumptions, we observed 377 of 16,976 phylogenetically confirmed HGT related to possible contamination in the host genome. The least contaminated genomes are those from the *Heterodera* genus, with no putative HGT assigned as contamination. In contrast, the *Meloidogyne hapla* shows the highest value with 139 proteins identified as contaminations (Fig. 2, supp. data table 2).

In addition, to be as stringent as possible, we deemed cases with less than three non-metazoan proteins in the sister branch as erroneously labelled in the reference library, and we then excluded 7,794 erroneous putative HGT, representing 11 to 32% of confirmed HGT depending on proteome size (Fig. 2, supp. data table 2).

After manually curating automated HGT detection, we finally identified a7,512 phylogenetically-confirmed HGT candidates corresponding to 1,209 acquisition events distributed in 592 trees. In addition, 1,113 putative HGT were classified as 'COMPLEX' related to 208 potential events spread on 131 trees. As we can notice, most of the "complex" cases (i.e. reduction from 9,411 to 1,209) were eliminated by the rule requiring more than three proteins in sister branches to declare a robust potential donor branch.

Therefore, on average, 0.2% of protein-coding genes of the *Tylenchina* genome stay qualified as 'complex' by automated processing, but 1.3% of genes have strong phylogenetic support for acquisition via horizontal gene transfers. The lower content of HGT candidates is identified in *B. xylophilus*, with only 96 HGT representing 55 HGT events (Fig. 2 A), spanning 0.5% of protein-coding genes of this genome (Supp data Fig. 1 A). In contrast, on average, we observe 1.38% of protein-coding genes that are HGT candidates in other PPNs. HGT candidates represent 1.23% and 1.47% of protein-coding genes in root-knot nematode and cyst nematode species, respectively.

Among cyst nematodes, the number of genes and events is homogenous (Fig. 2 A, Supp data, Fig. 1 A). In contrast, the number of HGT genes and events varies significantly between other species and particularly inside root-knot nematodes clade (Fig. 2 A). However, (except for *M. graminicola* where it's particularly high), the proportion of HGTs remains proportional to the number of coding proteins in PPN species (Supp data Fig. 1 A).

It should be noted that root-knot nematode genomes display substantial variations in genome size and gene number because of the polyploid nature of some of them. The genome presenting the highest proportion of genes acquired via HGT is the one of the burrowing nematode species *R. similis* (1.7%) (supplementary data Fig. 1). In addition, we observe between 0.10% (*H. glycines*) to 0.38% (*D. dipsaci*) of genes per species labelled 'complex' (Fig. 2).

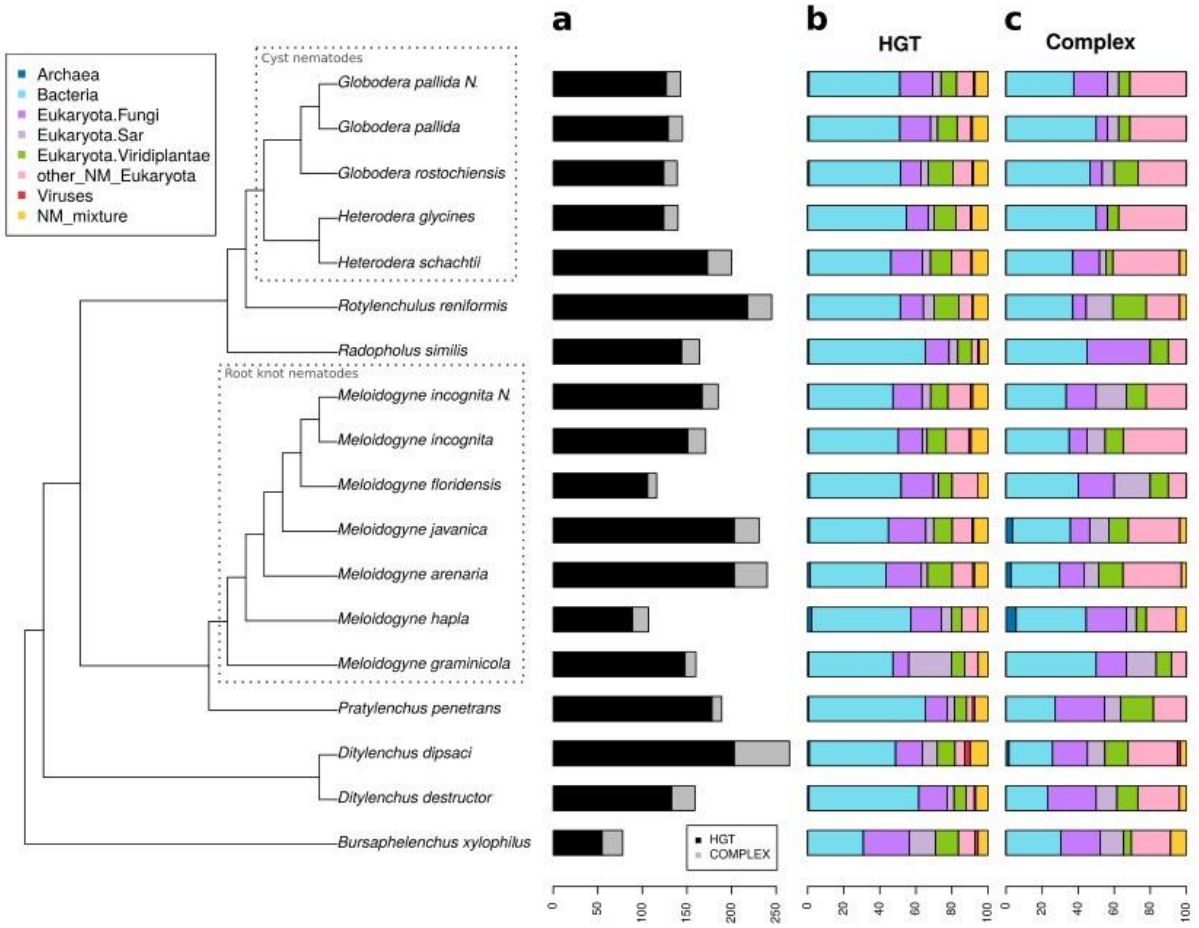


Figure 2 | The phylogenetic tree of studied species displaying (a) the classification of putative HGT events according to manually curated AvP results and (b) the kingdoms of potential donors of candidate HGT events, and (c) complex HGT events. The origins of donors were based on the sister branch taxa content (see main text)

3.3 Metagenomic data has highly expanded the catalogue of possible donors

3.3.1 Contribution of metagenomic data to the detection of HGT events

One of the main questions regarding horizontally acquired genes is the origin of these genes. Most cases of reported HGT in animal genomes seem to originate from bacterial genomes (Dunning Hotopp, 2011), but HGT from other taxa were also described in some animals from Viruses (Irwin et al., 2022), plants (Lapadula et al., 2020), Algae (Schwartz et al., 2014) and Fungi (Kikuchi et al., 2004). In PPN genomes, phylogenetically-assessed HGT candidates until today come from Bacterial or Fungal genomes presumed to live in the soil (Danchin et al., 2017, 2010; Haegeman et al., 2011; Kikuchi et al., 2004; Paganini et al., 2012). Populating our reference library with soil metagenomic data is expected to increase our knowledge on the origin of HGT candidates by including genes from the relatives and likely descendants of the original donor species.

The more comprehensive HGT detection in plant-parasitic nematode genomes supports the idea that at least some PPN genes originate from soil-dwelling microorganism genomes that were not represented in generalist sequence libraries. To assess how environmental data contribute to HGT

detection and how they complete the catalogue of potential donors, we analysed the content of *Tylenchina*'s HGT sister branches. All sources combined, the metagenomic data enriched the reconstructed phylogenetic trees with 171,779 proteins from uncultured microorganisms, 76,822 of which are found in the nearest sister branch of proteins classified as HGT or "complex". For 749 HGT acquisition events, corresponding to 5,966 HGT candidates (among which 130 complex events and 692 complex candidates), the nearest sister branch of *Tylenchina* proteins contains proteins from metagenomes. For 291 HGT events in *Tylenchina* genomes, the nearest sister branch contains more proteins from metagenomes than from NCBI nr. Several hundred HGT were detected only through the usage of metagenomic data: the nearest sister branch contains exclusively metagenomic data for 872 and 253 genes, respectively, classified as HGT and Complex (resp. 226 and 58 events).

3.3.2 Overview of the origins of horizontally acquired genes

To assess the origin of candidate HGTs, we automatically analysed the content of the closest sister branch to these genes in the phylogenetic tree. We deduce a donor kingdom if 80% of the taxa assigned to the proteins in the sister branch are from the same kingdom; otherwise, the origin is labelled "NM_mixture" for Non-Metazoan mixture.

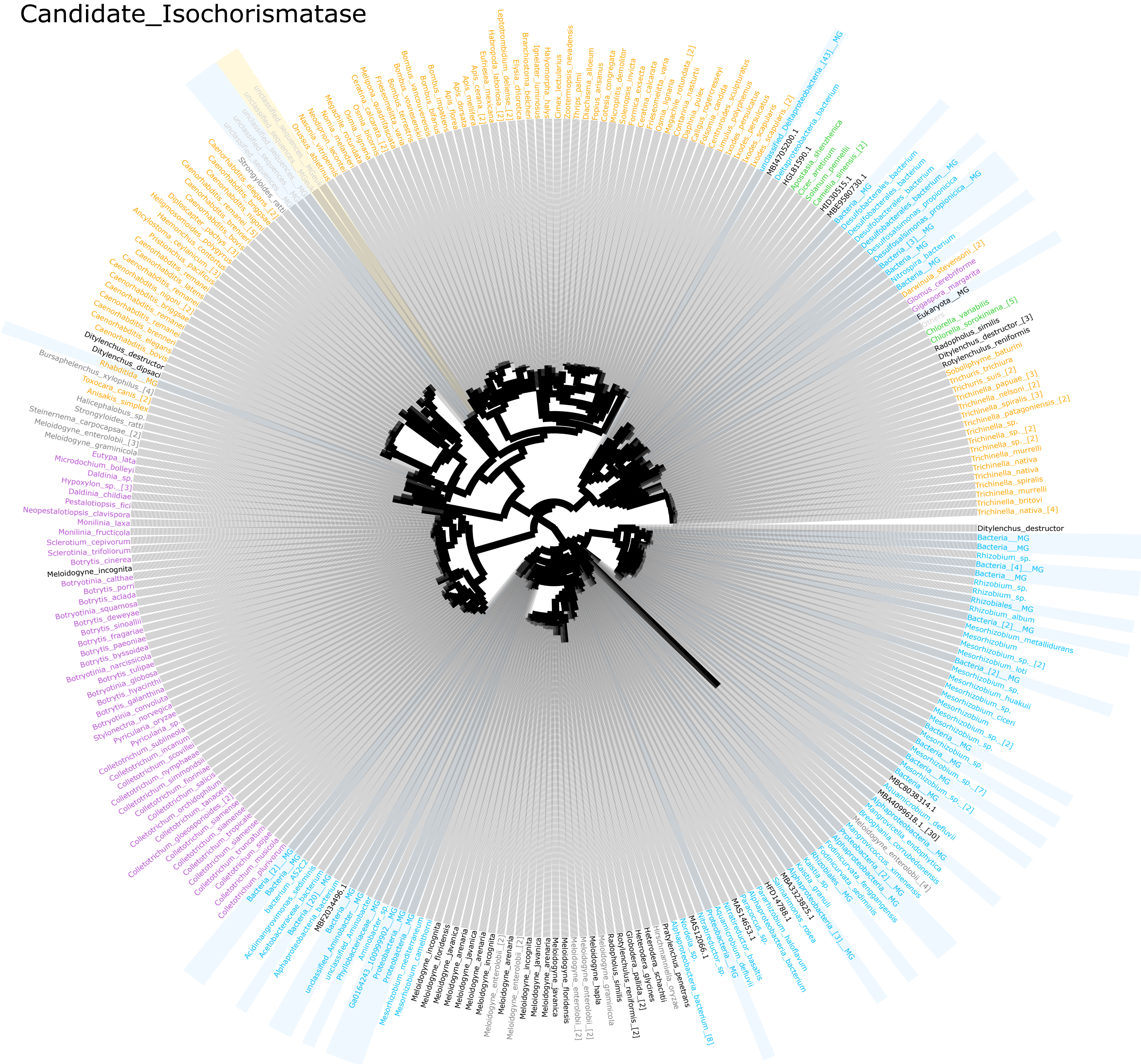
According to this donor identification process, bacteria is the most represented donor kingdom. From 55 (39% of HGT in *B. xylophilus*) to 623 (70% of HGT in R. similis) genes per genome seem to originate from this kingdom, originating from dozens of acquisition events. We observe 17 to 115 HGT events (respectively, *B. xylophilus* and *P. penetrans*) related to this kingdom corresponding to 55 and 214 genes. Hence, half of HGT in PPN genomes are of bacterial origin, except for *B. xylophilus*, where fungi are the majority (Fig. 2 B). In this species we observe 46 genes (47% of HGT) originating from 14 events of transfers from fungal donors. This is consistent with the dual lifestyle of *B. xylophilus*, which is able to feed both on plants and fungi. For other PPN species, fungal origin represents, on average, one-fifth of HGT events. At the scale of events, the smallest proportion of HGT events from fungi is observed in *M. hapla* (8.8% HGT) and the highest still in *B. xylophylus* genomes 33%. Besides bacteria and fungi, which had previously been described as HGT donors in PPN (Haegeman et al., 2011), we also identified other possible donors. For instance, several dozen genes of PPN appear to originate from various 'Protist' genomes with an average of 21 HGT per species (combining 'other_NM_Eukaryota' and 'Eukaryota@Sar' of Fig. 2 B). We observe the fewest HGT from this category in *M. floridensis*, with only 8.9% HGT, and the highest percentage of HGT in *D. dipsaci* with 26,6% (Fig. 2 B). Among those non-metazoan eukaryotes organisms, species from the SAR clade are the most represented for all PPN genomes studied. We observe, on average, ten genes per species from Archaea corresponding to between one and three events of transfers (Fig. 2 B). It is even less for Viruses, from which, on average, only one gene per species seems to have originated from (Fig. 2 B). Most of the events originating from Archaea and Viruses taxa are species-specific, and the representative proteins in sister branches are often in small numbers. Manual screening of phylogenetic trees revealed that no HGTs of viral origin appear to be strongly supported. In contrast, two events from Archaea origin shared by at least two and three species represent interesting new cases of HGT candidates. According to our HGT automated validation rules, for a few HGT candidates, the most likely donors are Viridiplantae species (Fig. 2 B). We observe between 8 (*B. xylophilus*) and 117 (*M. arenaria*) genes corresponding to 7 - 28 transfer events that seem to originate from plants. Satisfying our stringent validation rules, these genes are unlikely to represent contaminations. Nevertheless, we cannot define a single potential donor kingdom for a substantial proportion of genes between 5 to 10% according to nematode species due to a mixture of different non-metazoan proteins present in the sister branches with a proportion lower than 80%. As well as the HGT ratio in the

PPN genomes, the proportions of donor kingdoms are homogeneous among the PPNs of the superfamily *Tylenchoidea*. However, *B. xylophilus*, the sole representative of the *Aphelenchoidea* family shows a higher proportion of genes of fungal origin. Overall, the same proportion of donor kingdoms is observed for genes classified as complex.

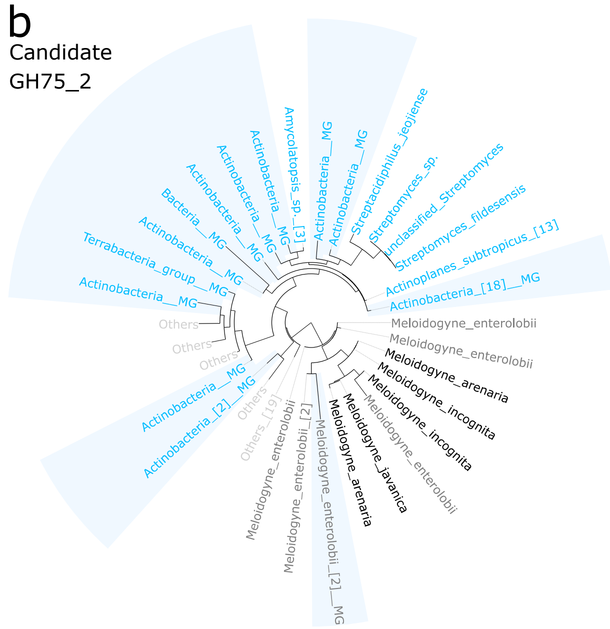
3.3.3 Improvement of gene donor identification

To assess more accurately donors' identities, we performed a taxonomic reduction with a last common ancestor on putative donor lineages. After extracting the complete lineage of PPN sister branches, we trimmed taxonomic rank until they encompassed 80%. For several hundred HGT events, we succeeded in defining a donor lineage at different taxonomic ranks shallower than a kingdom. Donor taxa were identified at the phylum rank for 63 HGT events displaying 17 different taxa. Half of the events seem to have originated from *the Proteobacteria* phylum including HGT coding for candidate Isochorismatase (Fig. 3 A). The *Actinobacteria* are the second most frequent donor phylum with 11 events from this phylum including HGT related to the GH75 family (Fig. 3 B).In other HGT phylogenetic trees, *Proteobacteria* donors could be more accurately identified as members of *Betaproteobacteria* class, such as HGT encoding polygalacturonases of GH28 family (Fig. 3 C). More rarely, *Proteobacteria* donors could also be identified at the family or genus level, such as for genes coding proteins involved in VB1, thiE and thiM synthesis (Fig. 3 D) that seems to originate from the *Burkholderiaceae* family. For 263 HGT events, putative donors could be more accurately identified at the species level (supplementary data fig. 3). However, a considerable part of the PPN sister branches contain proteins from distantly related species, and the taxonomic rank encompassing 80% of sister branch proteins is still the kingdom (supplementary data fig. 3). Consequently, for 434 and 293 events donor identities could only be identified respectively as Bacteria and Eukaryota. Re-predicted eukaryotic proteins from metagenomic data have also contributed to the catalogue of potential donors for HGT of PNN originating from eukaryotic donors. Multiple taxonomic levels of donor identification are also observed for these HGTs. For PPN genes acquired by horizontal transfer and encoding Thi4 proteins involved in vitamin B1 synthesis, the most comprehensive information regarding donor identity is the fungi kingdom as a whole due to the diversity of fungal donors in the sister branch (Fig. 4 A). In contrast, for a newly identified HGT in genomes of most endosedentary PNN genomes (RKN and CN), our analysis indicates these genes probably belong to the fungal class Agaricomycetes. The next closest sister branch originated from our re-predicted eukaryotic metagenomic dataset (Belliardo et al., 2022).

a
Candidate_Isochorismatase



b
Candidate
GH75_2



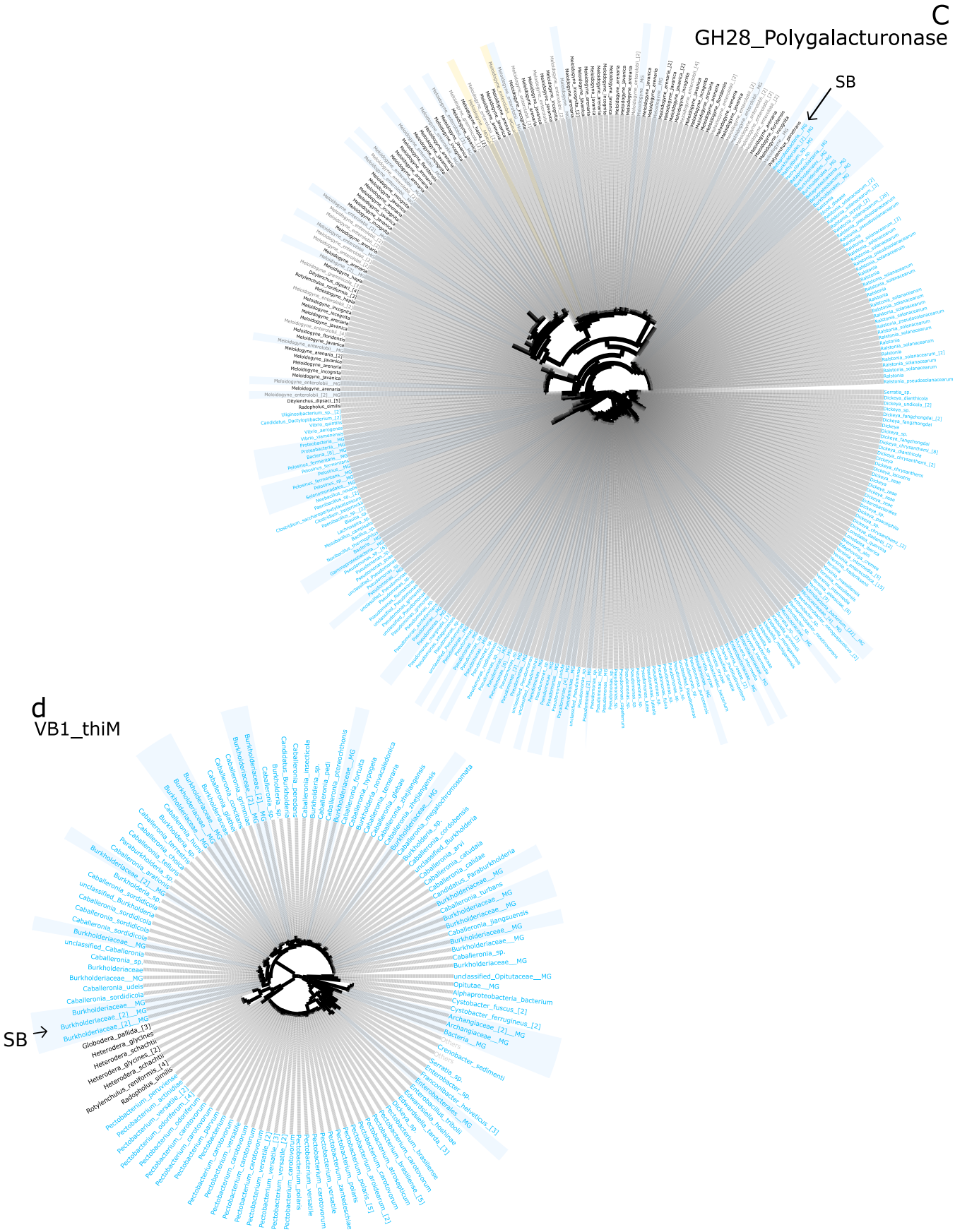


Figure 3 | The phylogenetic trees of four HGTs from bacterial donors in PPN genomes. Tree leaves are coloured according to the kingdom of taxa and data source (Bacteria: blue, Viridiplantae: green, Fungi: purple, Metazoa: Orange). Proteins from PPN are in black. Leaves highlighted in blue originated from public metagenomic data of IMGM and Mgnify (Chen et al., 2017; Mitchell et al., 2019), and those highlighted in yellow originated from eukaryotic proteins re-predicted from metagenomic data (Belliardo et al., 2022). Sister branch of PPN is indicated by 'SB'.

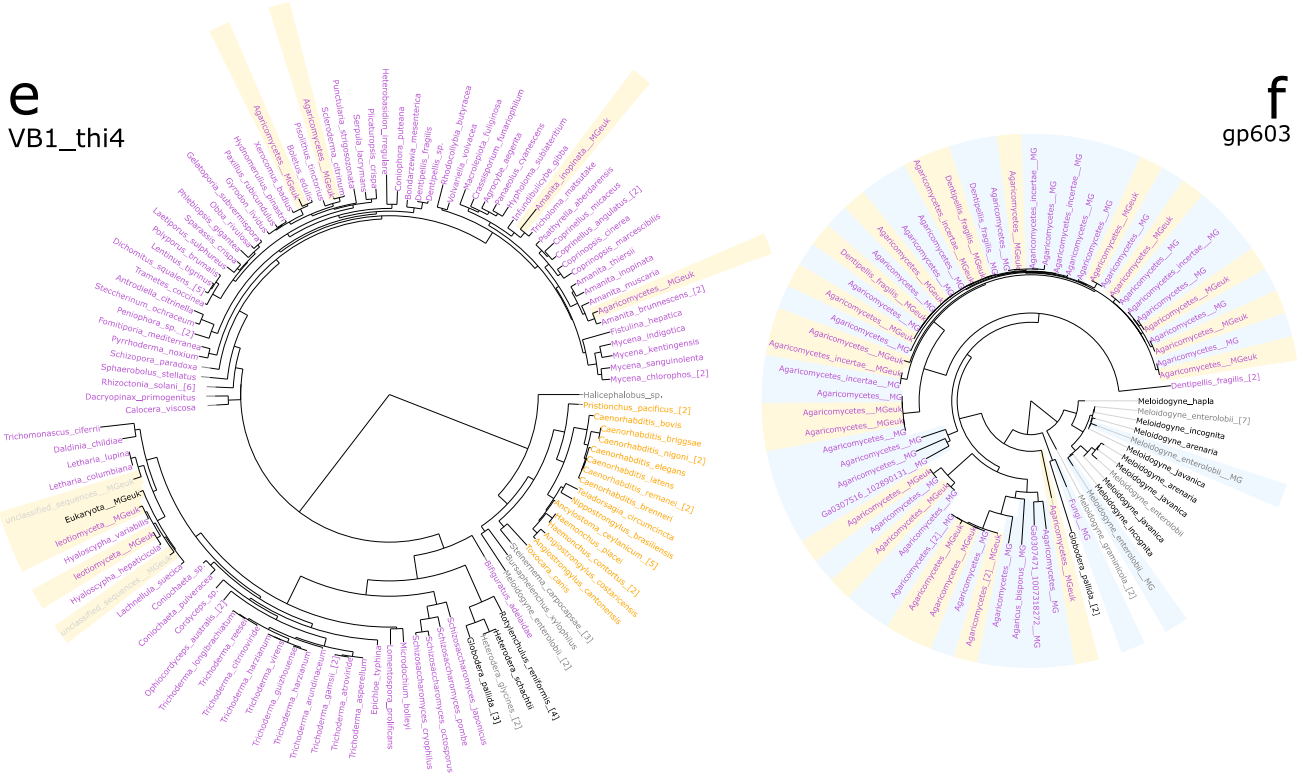


Figure 4 Phylogenetic trees of HGTs originating from fungal donors in PPN genomes. (A) The more accurate donor identity for HGT encoding thi4 VB1 is still at kingdom rank 'Fungi'. However, for (B) newly identified HGT, donors were identified at a shallower taxonomic level as Agaricomycetes fungal class. Tree leaves are coloured according to the kingdom of taxa and data source (PPN from proteome: black, PPN from the database: black, Fungi: purple, Metazoa: Orange). Non-PPN leaves are highlighted when they originated from metagenomic data in blue if the source is IMGM and Mgnify (Chen et al., 2017; Mitchell et al., 2019)and yellow whether they originated from eukaryotic proteins prediction from metagenomic data (Belliardo et al., 2022).

Furthermore, ancestral sister branches of HGTs are exclusively populated by proteins from metagenomes (Fig. 4 B). The presence of this gene in several nematode species and its localisation in contigs mainly composed of host genes further supports the hypothesis of integration of this fungal gene in the genomes of PNN. Moreover, the conservation of the gene in endoparasitic PPN species suggests that proteins coding by this HGT may play a role in the peculiar lifestyle of these nematodes. The only known conserved domain we could identify for these HGTs is a "CHRromatin Organisation MOdifier", this Pfam domain corresponding to 50 amino acids of the 350 AA constituting the protein.

3.4 The functional activity of HGT candidates

The use of metagenomic data in a large-scale analysis of PPN genomes revealed many cases of HGT, which could play substantial roles in the biology of these nematodes. To assess whether HGT candidates might have been involved in the evolutionary biology of *Tylenchina*, we performed a functional annotation followed by a Gene Ontology (GO) terms enrichment analysis.

Overall, across the 18 PPN genomes, at least one known protein domain, motif, or signal could be detected in 505,709 PPN proteins representing 77 to 93 % of predicted proteins depending on genomes. More specifically, at least one InterPro domain was identified for 378,821 of these proteins, which allowed the assignment of at least one GO term to 334,763 PPN proteins. In total, these labels represent 2,558 different GO terms, of which 633 are enriched among horizontally acquired genes.

3.4.1 Global description of HGT GO enrichment results

Thus, we searched significantly enriched GO terms in HGT candidates from the 18 studied PPN genomes. Considering all species, nearly half of HGT candidates had at least one GO term assigned (4,540/8,625), including 4,161 of the 7,512 PPN validated-HGT and 379 of the 1,113 complex HGT. Regarding only HGT candidates, we observed a total of 372 different enriched GO terms corresponding to 238 terms in the biological process (BP), 109 in the molecular function (MF) and 25 in the cellular component (CC) ontologies. The GoFuncR software identified 372 enriched GO terms in total, and 126 pruned terms in the refined version, excluding genes from significant child categories. We worked on the refined GO analysis results for more efficiency. Significantly enriched and refined functions of PPN HGTs covers 126 GO terms, including nine CCs, 56 MFs and 61 BPs associated with 3,254 HGTs (2,980, traceable HGTs and 274 Complex). We observe 78 enriched GO terms on average per PPN genomes (min: *M. graminicola*: 17 - max: *M. incognita Newton*: 174). Those results are consistent with genome size and HGT content of PPN genomes. Regarding cellular component terms, the 'extracellular region' appears significantly enriched in all PPN nematode clades except in *B. xylophilus,* representing *Aphelenchoides* taxa. In contrast, the terms "nucleus", "membrane", and "cytoplasm" are under-represented in horizontally acquired genes in PPN genomes. These cellular localisations suppose that proteins coded by HGTs act outside cells that produce them and could be effectors, small molecules secreted through the buccal stylet into the plant cell to modulate cell activity. Among biological process and molecular function GO terms, "carbohydrate metabolic process" and "hydrolase activity, hydrolysis of O-glycosyl compounds" are enriched in the HGT of all studied species. We also observed "Pectate lyase activity" molecular functions GO terms enriched in HGT of some PPN. These terms refer to the carbohydrate-degrading enzymes covering many biochemical activities such as cellulase, pectinase or arabinase activity (table 3) mainly active on sugars found in the plant cell wall. We also observed enriched GO terms related to vitamin B5 biosynthesis, such as "pantoate-beta-alanine ligase activity" molecular function and "pantothenate biosynthetic process" or "water-soluble vitamin biosynthetic process" in all Tylenchina species except in stem and bulbs nematode genomes. In the same species, we observed "cyanate metabolic process", "glutamine biosynthetic process" related to candidate cyanate lyases and candidate GSI glutamine synthase and "acetyltransferase activity" enriched GO terms relating to NodL-like. In contrast, candidate L-threonine aldolase proteins encoded by HGT were identified here in all Tylenchina species. Besides all the above-mentioned GO terms corresponding to previously described HGT events, we also identified other enriched GO terms probably representing so far undescribed HGT events. For instance, the GO terms "glucosylceramidase activity" or "isocitrate lyase activity" are also enriched in the HGT. For improved clarity, those terms were analysed in two separate batches; the first containing HGT related to the already described gene families, and the second HGT cases that showed potential new biological functions.

3.4.2 Distribution of HGT encoding gene families previously described from alien origins

Starting from previously described HGT (Abad et al., 2008; Bakker et al., 2001; Béra-Maillet et al., 2000; Danchin et al., 2017; Davis et al., 2000; Doyle and Lambert, 2002; Jaubert et al., 2002; Lambert et al., 1999; Ledger et al., 2006; McCarter et al., 2003; Mitreva-Dautova et al., 2006; Opperman et al., 2008; Paganini et al., 2012; Popeijus et al., 2000; Rosso et al., 1999; Scholl et al., 2003), some HGT encoding gene families historically known to originated bacteria or fungi are now labelled 'complex' in our analysis due to detection of homologous genes in other animal taxa.

Table 3 The distribution of gene families previously described as acquired by HGT in endo sedentary endoparasitic nematodes assessed by HGT detection in cyst nematodes (CN: G. pallida, G. pallida Newton, G. rostochiensis, H. glycines and H. schachtii), and root-knot nematodes (RKN: M. incognita, M. arenaria, M. javanica, M. hapla, M. graminicola), as well as in reniformis nematode (RN: R. reniformis), Burrow nematode (BN: R. similis), Lesion nematode (LN: Pratylenchus penetrans), Stem and bulb nematodes (SBN: D. destructor and D. dipsaci) and Pine wilt disease nematodes (PWD: B. xylophilus). Cutted donors' names mean Burkholderia.= Burkholderiaceae, Alphaprot. = Alphaproteobacteria and Betaproteobact. =Betaproteobacteria. For each gene family we provided HGT/events numbers.

Bio. Process	Molecular F.	Gene family	Donor	PPN_node	*CN*	*RN*	*BN*	*RKN*	*LN*	*SBN*	*PWN*
Plant cell wall degradation	Cellulose degradation	GH5_2 Cellulases	Bacteria	*Tylenchoidea Tylenchomorpha Tylenchina*	14/4	22/4	14/4	37/3	5/3	1/1	0
		GH45 Cellulases	Fungi	*Aphelenchoididae*	0	0	0	0	0	0	11/1
	Pectin decorations degradation	GH28 Polygalacturonase	**Betaproteobact.**	*Tylenchomorpha*	0	3/1	1/1	14/1	1/1	9/1	0
		GH43 candidate Arabinanase	Bacteria	*Tylenchoidea*	1/1	5/1	0	5/1	0	0	0
	Pectin degradation	PL3 Pectate Lyase	Bacteria	*Tylenchoidea, Meloidogyne*	5/1	0	0	19/3	4/1	0	0
	Pectinose / arabino. degradation	GH53 candidate Arabinogalactan galactosidase	Bacteria	*Tylenchoidea*	3/1	3/1	3/1	0	0	0	0
	Softening of non-covalent bonds	Expansin-like proteins	Bacteria	*Tylenchina*	8/1	9/1	4/1	32/1	1/1	6/1	6/1
	Xylan degradation	GH30 xylanase	Bacteria	*Meloidogyne*	0	0	0	19/1	0	0	0
Detoxification	Cyanide compounds detox.	Candidate Cyanate Lyases	Bacteria	*Tylenchoidea, Meloidogyne*	2/1	2/1	0	2/1	0	0	0
Feeding site induction	Candidate acetyltransferase	NodL - like	Bacteria	*Tylenchoidea*	1/1	9/1	3/1	2/1	4/1	0	0
Nutrient processing	Degradation of sucrose	GH32 invertase	Bacteria	*Tylenchomorpha*	10/1	12/1	4/1	6/1	6/1	2/1	0
	Nitrogen assimilation	Candidate GSI Glutamine Synthase	Bacteria	*Tylenchoidea*	2/1	4/1	5/1	6/1	2/1	0	0
	Vitamin B1 biosynthesis	VB1 tenA	Bacteria	*Tylenchomorpha*	2/1	12/1	0	0	0	3/1	0
		VB1 thi4	Fungi	*Tylenchoidea*	1/1	4/1	0	0	0	0	0
		VB1 thiD	Bacteria	*Tylenchoidea*	1/1	2/1	0	0	0	0	0
		VB1 thiE	**Burkholderiac.**	*Tylenchoidea*	1/1	2/1	1/1	0	0	0	0
		VB1 thiM	**Burkholderiac.**	*Tylenchoidea*	1/1	4/1	1/1	0	0	0	0
	Vitamin B5 biosynthesis	VB5 panC	Bacteria	*Tylenchoidea*	1/1	3/1	1/1	3/1	1/1	0	0
	Vitamin B6 biosynthesis	VB6 SOR-SNZ	Bacteria	*Tylenchomorpha*	1/1	0	1/1	0	1/1	1/1	0
		VB6 SNO	Bacteria	---	0	0	0	0	0	0	0
Plant defense manipulation	Conversion of Chorismate into SA	Candidate Isochorismatase	**Proteobacteria**	*Tylenchoidea*	1/1	2/1	1/1	3/1	1/1	0	0
		Chorismate Mutase	Bacteria	*Tylenchoidea*	2/2	4/2	2/1	3/1	1/1	0	0
Unknown	Unknown	Candidate L-threonine aldolase	Bacteria	*Tylenchomorpha*	4/1	5/1	1/1	3/1	1/1	1/1	0
		Candidate Phosphorybosyl transferase	**Alphaprot.**	*Tylenchomorpha*	3/1	5/1	2/1	3/1	1/1	1/1	0
		Candidate PolS Polyglutamate Synthase	Bacteria	*Tylenchoidea*	1/1	3/1	3/1	0	1/1	1/1	0

Almost all HGT coding for glycoside hydrolase family 5 enzymes with cellulase activity in the reference genome *M. incognita* were detected by our workflow but half of them were classified as "complex" because phytophagous insects also have cellulase coding genes from the same GH family (Busch et al., 2019; Pauchet et al., 2020; Shin et al., 2022). These results highlight that it is wiser to keep 'complex' genes in the subsequent analysis to prevent the over-exclusion of true horizontally acquired genes and reduce false negative HGT classification. Thus, our HGT detection successfully retrieved the gene families previously reported in the literature for PPN species. In addition, the integration of PNNs belonging to other nematode genera allowed us to clarify the distribution of HGT in PPN genomes and, thus, the history of these genes (Table 3).

Concerning Glycoside Hydrolases (GH), we observed HGT coding for proteins with cellulase activity belonging to the GH5 family and invertase activity belonging to the GH32 family in all PPN genomes from the *Tylenchoidea* family but not in *B. xylophilus*. This suggests these GH genes were acquired in a common ancestor of *Tylenchoidea*. Other cases suggest secondary loss after ancestral acquisition, such as, HGT encoding polygalacturonases of the GH28 family that was found in all *Tylenchoidea* species except cyst nematodes. Besides, HGT-encoding Expansin-like proteins are retrieved in all PPN species, including the Aphelenchoides representative. In contrast, HGT coding xylanase of family GH30 seem to be specific to root-knot nematodes so far, while candidate arabinogalactan of family GH53 seems specific to cyst nematodes and most closely related species (burrowing and reniform nematodes). We also retrieved candidate arabinase of a different family (GH43), that had so far been described exclusively in cyst and root-knot nematodes. For the first time, we observed that GH43 acquired by HGT are also present in the reniform nematode genome. Similarly, we observed pectate lyases from family PL3 widely conserved in PPN, including in root-knot, cyst and lesion nematode genomes but absent in burrowing and reniform nematodes.

Among the fifty HGT clusters containing genes from *B. xylophylus*, only ten include genes supporting enriched GO terms. From these ten events, only one is shared with others PPN (i.e. Expansin-like), and the nine other events are specific to *B. xylophilus*. We notice that, except HGT encoding Expansins-like proteins, gene families previously described in cyst and root-knot nematodes are clustered into a unique monophyletic clade, including only Tylenchoidea species. Among those nine HGT events, only two biochemical processes are found enriched in HGT regarding the rest of the genome, "carbohydrate metabolic process" and "protein phosphorylation". The "carbohydrate metabolic process" corresponds to only one HGT event specific to Aphelenchoides species, where putative donors are fungi from the Ascomycota phylum in our analysis, such as in (Kikuchi, 2004), although taxa diverge as species level.

Compared to the previously described HGT in *M. incognita*, we identified 426 new HGT corresponding to 139 acquisition events and 73 potential HGT labelled 'complex' from 20 possible acquisitions. These outcomes again give insight into the benefit of using metagenomic data in our reference library to expand the detection of HGT and the potential discovery of new HGT in *Tylenchina*. These results raise several questions: Are they also involved in the adaptative evolution of PPN? When were they acquired, and how did they evolve after acquisition in PPN genomes?

3.4.3 New functional domain detected in PPN HGT

First, we used a manual expert CaZyme annotation of the *M. incognita* proteome performed by the glycogenomics team to assess whether it is possible to identify previously undescribed CAZyme families and associated biological functions among the newly detected

HGTs. This examination allowed us to identify 27 *M. incognita* HGTs annotated as CaZymes "Glycoside Hydrolase Family 75" (GH75) and "GlycosylTransferase Family 4" (GT4). For each annotation, *M. incognita* HGT associated with the GH75 family were clustered in two different trees. GH75 is specific to root-knot nematodes, but GT4 is more widely distributed in all *Tylenchoidea* species except cyst nematodes. Concerning the putative functions of these CAZymes, so far the only function biologically validated in the GH75 family is chitosanase, so we can suggest a candidate chitosanase activity. In family GT4, in contrast, 24 different functions have been described (www.cazy.org) and inferring a possible function is not possible.

Then, to assess if, besides CAZymes, novel functions could be discovered in HGT newly identified in PPN genomes, we analysed GO enrichment in HGT candidates not related to previously described gene families. Several terms can be differentially associated with genes of the same HGT events or several events. However, a detailed analysis of term distribution across events indicates that in the majority of cases, the terms associated with the different HGTs within a same event are identical or similar. Also, when different terms are associated with the same gene, the same association is found for the majority of HGTs either completely or partially.

In the rest of this analysis, we focused on HGT events that were more likely to have substantially impacted the genome and biology of PPN. We selected HGT cases that were either conserved in at least three species and thus likely to have been retained during evolution after an ancestral acquisition, or that formed multigene families in at least one species, suggesting a possible evolutionary advantage of having multiple copies of the gene.Therefore, we filtered GO terms enriched in HGT of several PPN and found aggregates with the same terms in different genes. Thus, we identified 16 enriched molecular functions and height enriched biological processes found in varying numbers across PPN species (Table 5). No term commonly enriched in all studied PPN genus was identified among newly identified HGT. However, we can notice that all the enriched Molecular Function terms concerned various enzymatic activities. Some terms such as "3-methyl-2-oxobutanoate hydroxymethyltransferase activity" are detected in HGT widely distributed among nematodes.

In contrast, "dephospho-CoA kinase activity" is enriched in several PPN species but all being restricted to RKN. Overall, most of the enriched functions are associated with phylogenetic trees clustering several PPN genera of the *Tylencoidea* family. For seven biological processes and ten molecular functions, putative donors are Bacteria, including an accurate identification to the order level "*Spirochaetales*" for "thymidine kinase activity". Two molecular functions seem to have originated from Fungi, and one biological process from Viridiplantae, but for six terms, donor identity could not be assessed (Table 5).

Table 4 Candidate CaZy of in newly identified HGT

Bio. Process	Fonction Mol.	Famille de gènes	Donor	PPN_node	CN		RN	BN	RKN		LN	BSN	PWN
Unknown	Unknown	GH75 candidate chitosanase	Actinobacteria	*Tylenchomorpha*	N	0	0	0	N	2/2	0	0	0
		GT4 glycosyltransferase	Bacteria	*Meloidogyne*	N	0	1/1	2/1	N	2/2	3/1	2/1	0

Table 5 Enriched GO terms in newly detected HGT containing molecular functions (MF) and biological process (BP)

	Famille de gènes	*Donor*	*CN*	*RN*	*BN*	*RKN*	*LN*	*BSN*	*PWN*
MF	glucosylceramidase activity	Bacteria	N	N	Y	Y	Y	Y	Y
	serine-type endopeptidase activity	Bacteria/ Eukaryota@Sar	Y	N	N	Y	N	N	N
	cysteine-type peptidase activity	Fungi	N	Y	n	Y	N	Y	N
	glycine dehydrogenase (decarboxylating) activity	Bacteria	N	N	N	Y	N	N	N
	3-methyl-2-oxobutanoate hydroxymethyltransferase activity	Bacteria	N	Y	Y	Y	Y	Y	N
	polynucleotide adenylyltransferase activity	NM_mixture	Y	Y	N	Y	N	N	N
	thymidine kinase activity	Spirochaetales	Y	Y	Y	Y	Y	N	N
	isocitrate lyase activity	Bacteria	Y	N	Y	Y	N	Y	N
	malate synthase activity	Bacteria	Y	N	Y	Y	N	Y	N
	D-arabinono-1,4-lactone oxidase activity	Bacteria	Y	Y	Y	N	Y	Y	N
	nucleoside diphosphate kinase activity	Fungi	Y	Y	Y	N	Y	Y	N
	methyltransferase activity	Bacteria/Fungi	Y	Y	Y	Y	Y	Y	N
	xylulokinase activity	Bacteria	Y	Y	Y	N	Y	N	N
	acetyltransferase activity	Bacteria	Y	Y	Y	N	Y	N	N
	alanine-tRNA ligase activity	NM_mixture	N	Y	N	Y	N	N	N
	phosphatidylinositol phosphate kinase activity	NM_mixture	Y	Y	N	N	N	N	N
	dephospho-CoA kinase activity	Firmicutes	N	N	N	Y	N	N	N
BP	carbohydrate metabolic process; cellular aldehyde metabolic process	Bacteria	Y	N	Y	Y	N	Y	N
	purine ribonucleoside biosynthetic process	Actinobacteria /Fungi	Y	Y	N	Y	Y	N	N
	Hrd1p ubiquitin ligase complex	Bacteria	Y	Y	Y	Y	Y	Y	N
	transition metal ion homeostasis	Bacteria	Y	Y	Y	Y	Y	N	N
	sphingolipid metabolic process	Bacteria	N	Y	Y	Y	Y	Y	N
	RNA 3'-end processing; RNA polyadenylation	Viridiplantae	Y	Y	N	Y	N	N	N
	carbohydrate metabolic process; glyoxylate cycle	Bacteria	Y	N	Y	Y	N	Y	N
	telomere maintenance	Bacteria	Y	Y	N	Y	N	Y	N

Then, to assess if, besides CAZymes, novel functions could be discovered in HGT newly identified in PPN genomes, we analysed GO enrichment in HGT candidates not related to previously described gene families. Several terms can be differentially associated with genes of the same HGT events or several events. However, a detailed analysis of term distribution across events indicates that in the majority of cases, the terms associated with the different HGTs within a same event are identical or similar. Also, when different terms are associated with the same gene, the same association is found for the majority of HGTs either completely or partially.

In the rest of this analysis, we focused on HGT events that were more likely to have substantially impacted the genome and biology of PPN. We selected HGT cases that were either conserved in at least three species and thus likely to have been retained during evolution after an ancestral acquisition, or that formed

multigene families in at least one species, suggesting a possible evolutionary advantage of having multiple copies of the gene. Therefore, we filtered GO terms enriched in HGT of several PPN and found aggregates with the same terms in different genes. Thus, we identified 16 enriched molecular functions and height enriched biological processes found in varying numbers across PPN species (Table 5). No term commonly enriched in all studied PPN genus was identified among newly identified HGT. However, we can notice that all the enriched Molecular Function terms concerned various enzymatic activities. Some terms such as "3-methyl-2-oxobutanoate hydroxymethyltransferase activity" are detected in HGT widely distributed among nematodes. In contrast, "dephospho-CoA kinase activity" is enriched in several PPN species but all being restricted to RKN. Overall, most of the enriched functions are associated with phylogenetic trees clustering several PPN genera of the *Tylencoidea* family. For seven biological processes and ten molecular functions, putative donors are Bacteria, including an accurate identification to the order level "*Spirochaetales*" for "thymidine kinase activity". Two molecular functions seem clearly originating from Fungi, and one biological process from Viridiplantae. For six GO terms, donor identity could not be assessed (Table 5).

4. Discussion & Conclusion

4.1 Extension of the HGT detection in Tylenchina genomes

To assess the contribution of soil microorganism genes to PPN genomes, we performed an HGT screening using a reference library representative of the natural environment of these nematodes. We integrated 16 species covering different types of plant-parasitic lifestyles for a comprehensive understanding of this phenomenon in genetic novelty acquisition in PPN.

Using our soil-enriched protein library, we have detected between 96 to 1,090 genes validated HGT per PPN genome (supplementary data table 2). Regarding the number of genes, there is a high level of variation in HGT number between species, and even between species of the same genus (supplementary data).

For example, within the genus *Meloidogyne* itself, the different species have different ploidy levels and different haplotype resolution: some species are diploid and are assembled at a haploid consensus state due to low heterozygosity. While other species are polyploid (up to 4n) with most of the genome copies assembled separately, thanks to their high nucleotide divergence. Consequently, polyploid genomes tend to be much bigger and encode many more genes (up to 10 times more than diploid), the majority of which are duplicated. Moreover, a gene can undergo duplication events, after integration, which can vary in rate from one species to another depending on the dynamics of the genome. Thus, it could be more informative to compare acquisition events and percentage of the genes acquired by HGT.

Based on the hypothesis that three proteins constitute a robust enough sister branch, an HGT event was defined by a monophyletic cluster of PPN species followed by at least three non-metazoan proteins. According to this definition of HGT event, we observed more homogeneity between species in terms of number of events than comparison in gene number, such as for species of the *Globodera* genus (Fig. 2). We can note that a higher number of HGT events was reported in *R. reniformis*. Although HGT percentage in predicted proteins is substantially higher in the relative sister group of cyst nematodes than other genera, more exploration is required in *R. reniformis* to understand this specificity and its biological impact. The results of the HGT analysis integrating soil metagenomic data are consistent with previous reports of HGT in PPN genomes (Danchin et al., 2010b; Kikuchi et al., 2004). Nevertheless, between *Meloidogyne* species some variations among the number of events remain (Fig. 2).

The HGT ratio according to predicted proteins is quite homogenous between species, including *Meloidogyne* (supplementarity data fig. 1). This could suggest that a majority of

HGTs were acquired before the speciation of the *Tylenchina* clade but not only because of many HGT species-specific events. Indeed, on average, 50 HGT events per species seem to be species-specific with big variations as only 10 species-specific events were observed in *M. incognita* while 189 were observed in *D. dipsaci*. Part of the explanation may lie in sampling differences between the *Meloidogyne* genus (seven genomes for six different species) and the *Ditlylenchus* genus (2 genomes for 2 different species). Furthermore, the *Ditylenchus* genus holds an outgroup position relative to the rest of *Tylenchoidea* and thus orthologs of HGT in these species are less likely to be discovered. Except for *M. graminicola* displaying a high rate of species specific HGT events (77 events and 160 genes), this type of specific HGT is less common and corresponds to only dozens of genes/events. Another possibility is that, rather than representing species-specific acquisitions, these lineage-specific cases could represent older acquisitions followed by multiple convergent losses in the other lineages according to the results of Lai and al (Lai et al., 2022).

Finally, many HGT studies deployed on whole genomes estimate the proportion of genes acquired by HGT in PPN genomes to be between 2 and 3% (Lai et al., 2022; Phan et al., 2020). Here, after phylogenetic validation, we can reliably assume that about 1% of genes originate from horizontal transfer in plant-parasitic nematodes of the *Tylenchoidea* and 0.5% in *B. xylophylus*. Our analysis aimed at stringency and reliability, which may explain our lower estimate. Indeed this difference is probably due to the fact that previous estimates have been based on similarity scores with many potential false positives, including those due to taxonomic annotation errors. Phylogenetic analysis will invalidate these cases and return a more reliable yet conservative estimation of the percentage of genes acquired via HGT.

4.2 Comprehensive understanding of the timing of acquisition of HGT coding for gene families previously described

While previous analyses by Vanholme *et al.* claimed GH53 genes were specific to the *Heteroderea* nematode family (Vanholme et al., 2009), here we reported for the first time homologous HGT in *Rotylenchus reniformis* and *Radopholus similis* genomes. These two species have outgroup positions relative to the cyst nematode of the genus *Heterodera* and the presence of GH53 genes in their genomes suggest a more ancestral acquisition. And, indeed, according to corresponding tree topology (supp. data fig. 2), the most likely hypothesis is that this gene family originated in PPN from a single acquisition event of a bacterial gene in a common ancestor of cyst nematodes and *Radopholus*.

Similarly, we also reported for the first time HGT encoding VB1 thi4 proteins in *Rotylenchulus reniformis* genomes (Table 3, Fig. 4 A), and the phylogeny of orthologous HGT following the tree of PPN supports the hypothesis of a unique acquisition event of this fungal gene in a common ancestor of CN and reniform nematodes. Homologous HGTs are aggregated by species in phylogenetic trees for both gene families, suggesting that duplication events have occurred and most likely took place after speciation events.

Similarly, the HGTs encoding proteins with polygalacturonase activity from the GH28 family were so far described as RKN specific, but here we find GH28 genes acquired by HGT in all the *Tylenchidae* studied except cyst nematodes. This observation indicates that these genes were probably acquired in the common ancestor of the *Tylenchidea* and then lost specifically in the cyst nematodes. It would be interesting to investigate whether the dispensability of GH28 in cyst nematodes is due to a difference in the nature of the plants parasitized, the mode of parasitism or whether another gene plays a similar role in these species. This observation is consistent with multiple losses recently reported between different PPN species (Lai et al., 2022).

Furthermore gene copy number variations involving duplications and losses have also been observed between different populations within a same species in the PPN species *M. incognita* (Castagnone-Sereno et al., 2019).

Indeed, large-scale HGT detection in both clades 10 and 12 suggests a burst of acquisition events, duplication and losses in *Tyloichoidea* lineages. Most *B. xylophylus* HGT are specific to this species suggesting independent HGT acquisitions in *Aphelenchoides* and *Tyloichoidea* lineages, although HGT encoding Expansin-like protein clusters include *B. xylophylu*s. To further determine whether most HGT events are independent in *Aphelenchoides* and *Tyloichoidea* it will be necessary in the future to incorporate more members of the *Aphelenchoides* and also species having intermediate positions between these two groups such as *Aphelenchus avenae*. Indeed, in our analysis, *B. xylophilus* was the sole representent of the *Aphelenchoides* group and inclusion of other species from this group might reveal more shared HGT between the two groups. An interesting example is the GH5 family, which we retrieved in all PPN studies except the sole representative of *Aphelenchoidea* family in this study, *B. xylophilus*, suggesting that acquisition occurred in the last common ancestor of *Tylenchoidea*. However, Lai et al. recently identified GH5s with candidate cellulase activity in the genomes of *Aphelenchoides peseudobesseyi* and *Aphelenchoides bicaudatus*. These cellulases constitute a monophyletic group in their phylogenetic reconstructions, suggesting an acquisition in a common ancestor of *Tylenchoidea* and *Aphenlenchoides* followed by loss events in *Bursaphelenchus* (Lai et al., 2022). In the case of the PL3s, Lai and colleagues describe PL3s in Bursaphelenchus acquired by HGT of bacterial origin but from an acquisition event independent from the rest of the PPN. However, our analyses do not confirm these results. Although we identified PL3s from *Bursaphelenchus* in our phylogenies, they are not classified as HGTs for this species. In our analyses, all PPN genes potentially coding for PL3, including those of Bursaphelenchus were clustered in the same tree thanks to our meta-clustering methods. In this tree, proteins from *B. xylophylus* constitute an independent monophyletic group surrounded by metazoan proteins, resulting in a non-HGT classification. Therefore, it is unclear whether *Bursaphelenchus* have also acquired PL3 via an independent event.

Regardless of the limitations concerning the underrepresentation of Aphelenchoides in our analysis, many of the HGT identified in the *B. xylophilus* genome are found in specific phylogenetic trees containing no other PPN proteins, suggesting they originate from specific acquisition events. The alternative hypothesis is that those genes were acquired in a common ancestor of Aphelenchoides and *Tyloichoidea* but secondarily lost in *Tylenchoidea*. In other cases, *Tylenchoidea* equivalents of Bursaphelenchus proteins exist in the same tree but, in most cases, *Tylenchoidea* and *Aphelenchoides* are independently clustered with different microorganism lineages. These topologies suggest two independent acquisitions from different microorganisms.

Genes coding candidate Isochorismatase were classified as HGT in all *Tylenchoidea* species ranging from cyst nematodes to lesion nematodes (table 3), suggesting an ancient acquisition in a common ancestor. Supporting this possibility, homologous genes of those PPN are clustered into a monophyletic group following the phylogeny of this species (Figure 2 A). In the tree, this phylogenetic group of PPN is surrounded by many bacterial proteins, including numerous uncultured bacteria from metagenomic data. Besides these PPN, species candidate Isochorismatase genes were also identified in the stem and bulb nematodes from the *Ditylenchus* genus, which constitute an outgroup of the rest of these PNN. However, none of the candidate Isochorismatases from *Ditylenchus* were classified as HGT. Indeed, most *Ditylenchus* proteins fall within another branch, constituting a monophyletic group which includes both non-plant parasitic nematodes and arthropods. One hypothesis is that the gene found in *Ditylenchus* was inherited from a common ancestor of nematodes and

arthropods while the gene found in the other PPN was acquired via HGT of bacterial origin with this newly acquired gene replacing the original one inherited from a common ancestor.

4.3 Newly identified HGT

Our study provided some evidence that other essential biological functions could originate from HGT. If an acquired gene-coding DNA chunk provides no advantageous features to the host, we could hypothesise that the accumulation of deleterious mutations will not be counter-selected and that the gene will eventually become a pseudogene or be lost. On the other hand, if this DNA fragment provides an evolutionary advantage to its host, deleterious mutations will be counter-selected and the individuals possessing the gene will be eventually fixed at the population then species level. Indeed, individuals having one or several functional copies of this gene will be favoured by selection.

Previous studies have shown that HGTs involved in biological functions providing an evolutionary advantage in these parasites have undergone modification allowing the adaptation of foreign DNA to the genetic specificity of the host cell such as the acquisition of introns in the genes of bacterial origin, but also a conservation of functional domains and many duplications that may be involved in phenotypic plasticity. Similar evolutionary patterns were observed in newly detected HGT, suggesting those genes could play an important role in the biology of these parasites. This might be the case for the most enriched molecular function "xylulokinase activity" in HGT newly detected in all *Tylenchoidea* except in species of the *Meloidogyne* and *Ditylenchus* genus. It was demonstrated in yeast that xylulokinase catalyse D-xylulose processing, a major component of hemicellulose which is involved in the bridging function between the cellulose fibres, but also with other plant cell wall matrix compounds (Richard et al., 2000). Those genes may complete the list of enzymatic arsenal available to nematodes to degrade plant cell walls.

It would be interesting to isolate these genes in PPN species to perform a biochemical and functional characterization of the gene products.

All proteins in a tree are assumed to be homologous because they have been grouped based on strong similarity or occurrence of the same functional domain. For HGT originating from a single acquisition event, it is expected that PPNs would form a single monophyletic group following the species phylogeny. Consistent with this statement, in most cases, we observed only one PPN cluster, but several monophyletic PPN groups were found in the same tree. This observation is unlikely to result from an abusive grouping of our method because this phenomenon is also found for HGT coding for manually annotated PL3 or GH5 in the *M. incognita* genome. Indeed, these proteins are dispersed in several monophyletic PNN clusters, sometimes surrounded by highly different sister branches.

Actually, we do not know what happens after gene acquisition and even more when genes undergo numerous duplications, which can open a path to potential functional diversification. HGT encoding polygalacturonases have been identified in many species of herbivorous beetles. Kirsch *et al.* have shown that these enzymes originate from a single fungal gene acquisition event. The heterologous expression of about 50 of these genes has revealed a functional diversification that followed the acquisition. Here, the biochemical and functional characterisation of HGT copies in PPN might give more information about the origin and evolution after acquisition but also about the impact of those genes in nematode biology (Kirsch et al., 2014).

4.4 HGT in PPN genomes mostly originated from soil-dwelling microorganisms

In previous analyses, the results of Danchin and collaborators suggested that HGTs encoding invertase of the GH32 family in PPN genomes originated from bacteria of the genus

Ralstonia belonging to the class Betaproteobacteria (Danchin, 2016). In our analysis, the potential donors of these genes are identified merely as belonging to the class *Betaproteobacteria*. In our results, we find the same lineage as in previous studies but with a deeper / parental taxonomic rank here. This phenomenon is observed more widely for many HGTs. It stems from the last common ancestor taxonomic annotation procedures of the metagenomic data and the procedure of identifying the sister branch based on the set of leaves of this branch. This procedure provides less precise but more reliable information. It would be possible to refine the identity of proteins from metagenomic data using taxonomic annotation approaches such as the GTdb toolkit (Chaumeil et al., 2020). However, these approaches require higher computation times and much better genomic quality data (genome with 70% completeness), which are far from the features of the public data used here (Belliardo et al., 2022). Finally, it is essential to keep in mind that this taxonomic information is based on genomes and current species identity found in contemporary soils that are probably different from the ancestral micro-organisms that were the source of transfers.

The use of metagenomic data in a large-scale analysis of PPN revealed many new cases of HGT. Those genes conserved in all or part of the genomes may play a role in the peculiar lifestyle of these nematodes. In many cases, metagenomic data completed the phylogenetic trees and enriched the composition of sister branches. Interestingly, the sister branches of several dozens of HGTs events are exclusively populated by proteins from metagenomes and constitute a group of homologous proteins with more significant similarity to the nematode sequences than proteins coming from generalist sequence libraries (Fig. 4 A). The only functional domain that we could identify for these HGTs is a "CHRromatin Organisation MOdifier" Pfam domain corresponding to 50 amino acids of the 350 AA constituting the protein. However, the presence of this gene in several nematode species and its localization in contigs mainly composed of host genes supports the hypothesis of the horizontal origin of these genes. However, only 1/7 of the protein size corresponds to this functional domain, suggesting that other so far uncharacterized functional domains might compose these proteins.

Currently, most of the identified HGT are involved in bypassing plant defence mechanisms by either degrading the plant cell wall or detoxifying xenobiotics plant compounds but it is known that sedentary endoparasitic nematodes are also able to manipulate host genetic functioning and cell development to form the feeding site.

To conclude, the large-scale HGT detection in PPN genomes using a soil enriched protein library confirms that at least some of these genes originated in the soil. The most represented taxa in potential donors are *Proteobacteria* and *Actinobacteria*. Numerous soil and plant-plant associated bacteria species are from this clade. Phylogenetic analyses have refuted many putative HGT cases inferred by similarity scores only, highlighting the importance of phylogenetic validation and the high proportion of false positives resulting from similarity scores. This effort allowed us to assess more accurately the real contribution of the soil-dwelling species to the genomes of these nematodes. We can now confirm that the proportion of genes acquired through this pathway represents about 1% of the protein coding genes. This large scale screening of whole proteomes from sixteen species gives a deeper understanding of the acquisition timing and distribution among species of historical HGT. Furthermore, the results of this study pave the way for further functional exploration to discover new biological functions acquired by the HGT in PPN genomes.

References

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M.-C., Coutinho, P.M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T.R., Markov, G.V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T.J., Blaxter, M., Bleve-Zacheo, T., Davis, E.L., Ewbank, J.J., Favery, B., Grenier, E., Henrissat, B., Jones, J.T., Laudet, V., Maule, A.G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J., Wincker, P., 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol. 26, 909–915. https://doi.org/10.1038/nbt.1482

Ahmed, M., Roberts, N.G., Adediran, F., Smythe, A.B., Kocot, K.M., Holovachov, O., 2022. Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. Front. Ecol. Evol. 9, 769565. https://doi.org/10.3389/fevo.2021.769565

Alačević, M., 1963. Interspecific Recombination in Streptomyces. Nature 197, 1323–1323. https://doi.org/10.1038/1971323a0

Arnold, B.J., Huang, I.-T., Hanage, W.P., 2022. Horizontal gene transfer and adaptive evolution in bacteria. Nat. Rev. Microbiol. 20, 206–218. https://doi.org/10.1038/s41579-021-00650-4

Bakker, J., Gommers, F., Smant, G., Abad, P., Rosso, M.-N., Dautova, M., 2001. Single pass cDNA sequencing - a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode Meloidogyne incognita. Nematology 3, 129–139. https://doi.org/10.1163/156854101750236259

Belliardo, C., Koutsovoulos, G.D., Rancurel, C., Clément, M., Lipuma, J., Bailly-Bechet, M., Danchin, E.G.J., 2022. Improvement of eukaryotic protein predictions from soil metagenomes. Sci. Data 9, 311. https://doi.org/10.1038/s41597-022-01420-4

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2012. GenBank. Nucleic Acids Res. 41, D36–D42. https://doi.org/10.1093/nar/gks1195

Béra-Maillet, C., Arthaud, L., Abad, P., Rosso, M.-N., 2000. Biochemical characterization of MI-ENG1, a family 5 endoglucanase secreted by the root-knot nematode Meloidogyne incognita: Root-knot nematode cellulase characterization. Eur. J. Biochem. 267, 3255–3263. https://doi.org/10.1046/j.1432-1327.2000.01356.x

Bird, D.McK., Jones, J.T., Opperman, C.H., Kikuchi, T., Danchin, E.G.J., 2015. Signatures of adaptation to plant parasitism in nematode genomes. Parasitology 142, S71–S84. https://doi.org/10.1017/S0031182013002163

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49, D344–D354. https://doi.org/10.1093/nar/gkaa977

Boto, L., 2014. Horizontal gene transfer in the

acquisition of novel traits by metazoans. Proc. R. Soc. B Biol. Sci. 281. https://doi.org/10.1098/rspb.2013.2450

Brent, M.R., 2007. How does eukaryotic gene prediction work? Nat. Biotechnol. 25, 883–885. https://doi.org/10.1038/nbt0807-883

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176

Busch, A., Danchin, E.G.J., Pauchet, Y., 2019. Functional diversification of horizontally acquired glycoside hydrolase family 45 (GH45) proteins in Phytophaga beetles. BMC Evol. Biol. 19. https://doi.org/10.1186/s12862-019-1429-9

Castagnone-Sereno, P., Mulet, K., Danchin, E.G.J., Koutsovoulos, G.D., Karaulic, M., Da Rocha, M., Bailly-Bechet, M., Pratx, L., Perfus-Barbeoch, L., Abad, P., 2019. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. Mol. Ecol. 28, 2559–2572. https://doi.org/10.1111/mec.15095

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, I.-M.A., Markowitz, V.M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., Varghese, N., Hadjithomas, M., Tennessen, K., Nielsen, T., Ivanova, N.N., Kyrpides, N.C., 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 45, D507–D516. https://doi.org/10.1093/nar/gkw929

Craig, J.P., Bekal, S., Niblack, T., Domier, L., Lambert, K.N., 2009. Evidence for Horizontally Transferred Genes Involved in the Biosynthesis of Vitamin B1, B5, and B7 in Heterodera glycines 10.

Danchin, E., Perfus-Barbeoch, L., Rancurel, C., Thorpe, P., Da Rocha, M., Bajew, S., Neilson, R., (Guzeeva), E.S., Da Silva, C., Guy, J., Labadie, K., Esmenjaud, D., Helder, J., Jones, J., den Akker, S., 2017. The Transcriptomes of Xiphinema index and Longidorus elongatus Suggest Independent Acquisition of Some Plant Parasitism Genes by Horizontal Gene Transfer in Early-Branching Nematodes. Genes 8, 287. https://doi.org/10.3390/genes8100287

Danchin, E.G.J., 2016. Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube? BMC Biol. 14. https://doi.org/10.1186/s12915-016-0330-x

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., Abad, P., 2010a. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. 107, 17651–17656. https://doi.org/10.1073/pnas.1008486107

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., Abad, P., 2010b. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. 201008486. https://doi.org/10.1073/pnas.1008486107

Davis, E.L., Hussey, R.S., Baum, T.J., Bakker, J., Schots, A., Rosso, M.-N., Abad, P., 2000. Nematode Parasitism Genes. Annu. Rev. Phytopathol. 38, 365–396. https://doi.org/10.1146/annurev.phyto.38.1.365

Decraemer, Wilfrida and Hunt, David J, 2006. Structure and classification, in: Perry, Roland N and Moens, Maurice (Ed.), Plant Nematology. CABI, pp. 3–32.

Doyle, E.A., Lambert, K.N., 2002. Cloning and Characterization of an Esophageal-Gland-Specific Pectate

Lyase from the Root-Knot Nematode *Meloidogyne javanica*. Mol. Plant-Microbe Interactions® 15, 549–556. https://doi.org/10.1094/MPMI.2002.15.6.549

Dunning Hotopp, J.C., 2011. Horizontal gene transfer between bacteria and animals. Trends Genet. 27, 157–163. https://doi.org/10.1016/j.tig.2011.01.005

Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16. https://doi.org/10.1186/s13059-015-0721-2

Eves-van den Akker, S., Laetsch, D.R., Thorpe, P., Lilley, C.J., Danchin, E.G.J., Da Rocha, M., Rancurel, C., Holroyd, N.E., Cotton, J.A., Szitenberg, A., Grenier, E., Montarry, J., Mimee, B., Duceppe, M.-O., Boyes, I., Marvin, J.M.C., Jones, L.M., Yusup, H.B., Lafond-Lapalme, J., Esquibet, M., Sabeh, M., Rott, M., Overmars, H., Finkers-Tomczak, A., Smant, G., Koutsovoulos, G., Blok, V., Mantelin, S., Cock, P.J.A., Phillips, W., Henrissat, B., Urwin, P.E., Blaxter, M., Jones, J.T., 2016. The genome of the yellow potato cyst nematode, Globodera rostochiensis, reveals insights into the basis of parasitism and virulence. Genome Biol. 17, 124. https://doi.org/10.1186/s13059-016-0985-1

Fierer, N., 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579–590. https://doi.org/10.1038/nrmicro.2017.87

Gilbert, C., Belliardo, C., 2022. The diversity of endogenous viral elements in insects. Curr. Opin. Insect Sci. 49, 48–55. https://doi.org/10.1016/j.cois.2021.11.007

Gladyshev, E.A., Meselson, M., Arkhipova, I.R., 2008. Massive horizontal gene transfer in bdelloid rotifers. Science 320, 1210–1213. https://doi.org/10.1126/science.1156407

Grote, S., 2017. GOfuncR. https://doi.org/10.18129/B9.BIOC.GOFUNCR

Grynberg, P., Coiti Togawa, R., Dias de Freitas, L., Antonino, J.D., Rancurel, C., Mota do Carmo Costa, M., Grossi-de-Sa, M.F., Miller, R.N.G., Brasileiro, A.C.M., Messenberg Guimaraes, P., Danchin, E.G.J., 2020. Comparative Genomics Reveals Novel Target Genes towards Specific Control of Plant-Parasitic Nematodes. Genes 11, 1347. https://doi.org/10.3390/genes11111347

Haegeman, A., Jones, J.T., Danchin, E.G.J., 2011. Horizontal gene transfer in nematodes: a catalyst for plant parasitism? Mol. Plant-Microbe Interact. MPMI 24, 879–887. https://doi.org/10.1094/MPMI-03-11-0055

Holterman, M., Karegar, A., Mooijman, P., van Megen, H., van den Elsen, S., Vervoort, M.T.W., Quist, C.W., Karssen, G., Decraemer, W., Opperman, C.H., Bird, D.M., Kammenga, J., Goverse, A., Smant, G., Helder, J., 2017. Disparate gain and loss of parasitic abilities among nematode lineages. PLOS ONE 12, e0185445. https://doi.org/10.1371/journal.pone.0185445

Husnik, F., McCutcheon, J.P., 2018. Functional horizontal gene transfer from bacteria to eukaryotes. Nat. Rev. Microbiol. 16, 67–79. https://doi.org/10.1038/nrmicro.2017.137

Irwin, N.A.T., Pittis, A.A., Richards, T.A., Keeling, P.J., 2022. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat. Microbiol. 7, 327–336. https://doi.org/10.1038/s41564-021-01026-3

Jaubert, S., Laffaire, J.-B., Abad, P., Rosso, M.-N., 2002. A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne*

incognita [1]. FEBS Lett. 522, 109–112. https://doi.org/10.1016/S0014-5793(02)02906-X

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kikuchi, T., Jones, J.T., Aikawa, T., Kosaka, H., Ogura, N., 2004. A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus*. FEBS Lett. 572, 201–205. https://doi.org/10.1016/j.febslet.2004.07.039

Kirsch, R., Gramzow, L., Theißen, G., Siegfried, B.D., ffrench-Constant, R.H., Heckel, D.G., Pauchet, Y., 2014. Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. Insect Biochem. Mol. Biol. 52, 33–50. https://doi.org/10.1016/j.ibmb.2014.06.008

Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A., Blaxter, M., 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc. Natl. Acad. Sci. 113, 5053–5058. https://doi.org/10.1073/pnas.1600338113

Koutsovoulos, G.D., Granjeon Noriot, S., Bailly-Bechet, M., Danchin, E.G.J., Rancurel, C., 2022a. AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. PLOS Comput. Biol. 18, e1010686. https://doi.org/10.1371/journal.pcbi.1010686

Koutsovoulos, G.D., Granjeon Noriot, S., Bailly-Bechet, M., Danchin, E.G.J., Rancurel, C., 2022b. AvP: a software package for automatic phylogenetic detection of candidate horizontal gene transfers. (preprint). Bioinformatics. https://doi.org/10.1101/2022.06.23.497291

Lai, C.-K., Lee, Y., Ke, H.-M., Lu, M.R., Liu, W.-A., Lee, H.-H., Liu, Y.-C., Yoshiga, T., Kikuchi, T., Chen, P.J., Tsai, I.J., 2022. The *Aphelenchoides* genomes reveal major events of horizontal gene transfers in clade IV nematodes (preprint). Genomics. https://doi.org/10.1101/2022.09.13.507733

Lambert, K.N., Allen, K.D., Sussex, I.M., 1999. Cloning and Characterization of an Esophageal-Gland-Specific Chorismate Mutase from the Phytoparasitic Nematode *Meloidogyne javanica*. Mol. Plant-Microbe Interactions® 12, 328–336. https://doi.org/10.1094/MPMI.1999.12.4.328

Lapadula, W.J., Mascotti, M.L., Juri Ayub, M., 2020. Whitefly genomes contain ribotoxin coding genes acquired from plants. Sci. Rep. 10, 15503. https://doi.org/10.1038/s41598-020-72267-1

Ledger, T.N., Jaubert, S., Bosselut, N., Abad, P., Rosso, M.-N., 2006. Characterization of a new β-1,4-endoglucanase gene from the root-knot nematode Meloidogyne incognita and evolutionary scheme for phytonematode family 5 glycosyl hydrolases. Gene 382, 121–128. https://doi.org/10.1016/j.gene.2006.06.023

Leger, M.M., Eme, L., Stairs, C.W., Roger, A.J., 2018. Demystifying Eukaryote Lateral Gene Transfer (Response to Martin). BioEssays 40, 1700242. https://doi.org/10.1002/bies.201700242

Loiseau, V., Peccoud, J., Bouzar, C., Guillier, S., Fan, J., Gueli Alletti, G., Meignin, C., Herniou, E.A., Federici, B.A., Wennmann, J.T., Jehle, J.A., Cordaux, R., Gilbert, C., 2021. Monitoring Insect Transposable Elements in Large Double-Stranded DNA Viruses Reveals Host-to-Virus and Virus-to-Virus Transposition. Mol. Biol. Evol. 38, 3512–3530. https://doi.org/10.1093/molbev/msab198

Maule, A.G., Marks, N.J. (Eds.), 2006. Parasitic flatworms: molecular biology, biochemistry, immunology and physiology. CABI, Wallingford, UK ; Cambridge, MA.

McCarter, J.P., Dautova Mitreva, M., Martin, J., Dante, M., Wylie, T., Rao, U., Pape, D., Bowers, Y., Theising, B., Murphy, C.V., Kloek, A.P., Chiapelli, B.J., Clifton, S.W., Bird, D.M., Waterston, R.H., 2003. Analysis and functional classification of transcripts from the nematode Meloidogyne incognita. Genome Biol. 4, R26. https://doi.org/10.1186/gb-2003-4-4-r26

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., 2019. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. gkz1035. https://doi.org/10.1093/nar/gkz1035

Mitreva-Dautova, M., Roze, E., Overmars, H., de Graaff, L., Schots, A., Helder, J., Goverse, A., Bakker, J., Smant, G., 2006. A Symbiont-Independent Endo-1,4-β-Xylanase from the Plant-Parasitic Nematode *Meloidogyne incognita*. Mol. Plant-Microbe Interactions® 19, 521–529. https://doi.org/10.1094/MPMI-19-0521

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T.B.K., Nielsen, T., Kirton, E., Faria, J.P., Edirisinghe, J.N., Henry, C.S., Jungbluth, S.P., Chivian, D., Dehal, P., Wood-Charlson, E.M., Arkin, A.P., Tringe, S.G., Visel, A., IMG/M Data Consortium, Abreu, H., Acinas, S.G., Allen, E., Allen, M.A., Alteio, L.V., Andersen, G., Anesio, A.M., Attwood, G., Avila-Magaña, V., Badis, Y., Bailey, J., Baker, B., Baldrian, P., Barton, H.A., Beck, D.A.C., Becraft, E.D., Beller, H.R., Beman, J.M., Bernier-Latmani, R., Berry, T.D., Bertagnolli, A., Bertilsson, S., Bhatnagar, J.M., Bird, J.T., Blanchard, J.L., Blumer-Schuette, S.E., Bohannan, B., Borton, M.A., Brady, A., Brawley, S.H., Brodie, J., Brown, S., Brum, J.R., Brune, A., Bryant, D.A., Buchan, A., Buckley, D.H., Buongiorno, J., Cadillo-Quiroz, H., Caffrey, S.M., Campbell, A.N., Campbell, B., Carr, S., Carroll, J., Cary, S.C., Cates, A.M., Cattolico, R.A., Cavicchioli, R., Chistoserdova, L., Coleman, M.L., Constant, P., Conway, J.M., Mac Cormack, W.P., Crowe, S., Crump, B., Currie, C., Daly, R., DeAngelis, K.M., Denef, V., Denman, S.E., Desta, A., Dionisi, H., Dodsworth, J., Dombrowski, N., Donohue, T., Dopson, M., Driscoll, T., Dunfield, P., Dupont, C.L., Dynarski, K.A., Edgcomb, V., Edwards, E.A., Elshahed, M.S., Figueroa, I., Flood, B., Fortney, N., Fortunato, C.S., Francis, C., Gachon, C.M.M., Garcia, S.L., Gazitua, M.C., Gentry, T., Gerwick, L., Gharechahi, J., Girguis, P., Gladden, J., Gradoville, M., Grasby, S.E., Gravuer, K., Grettenberger, C.L., Gruninger, R.J., Guo, J., Habteselassie, M.Y., Hallam, S.J., Hatzenpichler, R., Hausmann, B., Hazen, T.C., Hedlund, B., Henny, C., Herfort, L., Hernandez, M., Hershey, O.S., Hess, M., Hollister, E.B., Hug, L.A., Hunt, D., Jansson, J., Jarett, J., Kadnikov, V.V., Kelly, C., Kelly, R., Kelly, W.,

Kerfeld, C.A., Kimbrel, J., Klassen, J.L., Konstantinidis, K.T., Lee, L.L., Li, W.-J., Loder, A.J., Loy, A., Lozada, M., MacGregor, B., Magnabosco, C., Maria da Silva, A., McKay, R.M., McMahon, K., McSweeney, C.S., Medina, M., Meredith, L., Mizzi, J., Mock, T., Momper, L., Moran, M.A., Morgan-Lang, C., Moser, D., Muyzer, G., Myrold, D., Nash, M., Nesbø, C.L., Neumann, A.P., Neumann, R.B., Noguera, D., Northen, T., Norton, J., Nowinski, B., Nüsslein, K., O'Malley, M.A., Oliveira, R.S., Maia de Oliveira, V., Onstott, T., Osvatic, J., Ouyang, Y., Pachiadaki, M., Parnell, J., Partida-Martinez, L.P., Peay, K.G., Pelletier, D., Peng, X., Pester, M., Pett-Ridge, J., Peura, S., Pjevac, P., Plominsky, A.M., Poehlein, A., Pope, P.B., Ravin, N., Redmond, M.C., Reiss, R., Rich, V., Rinke, C., Rodrigues, J.L.M., Rodriguez-Reillo, W., Rossmassler, K., Sackett, J., Salekdeh, G.H., Saleska, S., Scarborough, M., Schachtman, D., Schadt, C.W., Schrenk, M., Sczyrba, A., Sengupta, A., Setubal, J.C., Shade, A., Sharp, C., Sherman, D.H., Shubenkova, O.V., Sierra-Garcia, I.N., Simister, R., Simon, H., Sjöling, S., Slonczewski, J., Correa de Souza, R.S., Spear, J.R., Stegen, J.C., Stepanauskas, R., Stewart, F., Suen, G., Sullivan, M., Sumner, D., Swan, B.K., Swingley, W., Tarn, J., Taylor, G.T., Teeling, H., Tekere, M., Teske, A., Thomas, T., Thrash, C., Tiedje, J., Ting, C.S., Tully, B., Tyson, G., Ulloa, O., Valentine, D.L., Van Goethem, M.W., VanderGheynst, J., Verbeke, T.J., Vollmers, J., Vuillemin, A., Waldo, N.B., Walsh, D.A., Weimer, B.C., Whitman, T., van der Wielen, P., Wilkins, M., Williams, T.J., Woodcroft, B., Woolet, J., Wrighton, K., Ye, J., Young, E.B., Youssef, N.H., Yu, F.B., Zemskaya, T.I., Ziels, R., Woyke, T., Mouncey, N.J., Ivanova, N.N., Kyrpides, N.C., Eloe-Fadrosh, E.A., 2021. A genomic catalog of Earth's microbiomes. Nat. Biotechnol. 39, 499–509. https://doi.org/10.1038/s41587-020-0718-6

Naylor, D., Fansler, S., Brislawn, C., Nelson, W.C., Hofmockel, K.S., Jansson, J.K., McClure, R., 2020. Deconstructing the Soil Microbiome into Reduced-Complexity Functional Modules. mBio 11. https://doi.org/10.1128/mBio.01349-20

Noon, J.B., Baum, T.J., 2016. Horizontal gene transfer of acetyltransferases, invertases and chorismate mutases from different bacteria to diverse recipients. BMC Evol. Biol. 16, 74. https://doi.org/10.1186/s12862-016-0651-y

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P., Windham, E., 2008. Sequence and genetic map of

Meloidogyne hapla : A compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. 105, 14802–14807. https://doi.org/10.1073/pnas.0805946105

Paez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., Huntemann, M., K.?Reddy, T.B., Pavlopoulos, G.A., Sullivan, M.B., Campbell, B.J., Chen, F., McMahon, K., Hallam, S.J., Denef, V., Cavicchioli, R., Caffrey, S.M., Streit, W.R., Webster, J., Handley, K.M., Salekdeh, G.H., Tsesmetzis, N., Setubal, J.C., Pope, P.B., Liu, W.-T., Rivers, A.R., Ivanova, N.N., Kyrpides, N.C., 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 45, D457–D465. https://doi.org/10.1093/nar/gkw1030

Paganini, J., Campan-Fournier, A., Da Rocha, M., Gouret, P., Pontarotti, P., Wajnberg, E., Abad, P., Danchin, E.G.J., 2012. Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. PLoS ONE 7, e50875. https://doi.org/10.1371/journal.pone.0050875

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W., 2018. Author Correction: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 3, 253–253. https://doi.org/10.1038/s41564-017-0083-5

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pauchet, Y., Ruprecht, C., Pfrengle, F., 2020. Analyzing the Substrate Specificity of a Class of Long-Horned-Beetle-Derived Xylanases by Using Synthetic Arabinoxylan Oligo- and Polysaccharides. Chembiochem Eur. J. Chem. Biol. 21, 1517–1525. https://doi.org/10.1002/cbic.201900687

Phan, N.T., Orjuela, J., Danchin, E.G.J., Klopp, C., Perfus-Barbeoch, L., Kozlowski, D.K., Koutsovoulos, G.D., Lopez-Roques, C., Bouchez, O., Zahm, M., Besnard, G., Bellafiore, S., 2020. Genome structure and content of the rice root-knot nematode (*Meloidogyne graminicola*). Ecol. Evol. 10, 11006–11021. https://doi.org/10.1002/ece3.6680

Pienaar, R.D., Gilbert, C., Belliardo, C., Herrero, S., Herniou, E.A., 2022. First Evidence of Past and Present Interactions between Viruses and the Black Soldier Fly, Hermetia illucens. Viruses 14, 1274. https://doi.org/10.3390/v14061274

Popeijus, H., Overmars, H., Jones, J., Blok, V., Goverse, A., Helder, J., Schots, A., Bakker, J., Smant, G., 2000. Degradation of plant cell walls by a nematode. Nature 406, 36–37. https://doi.org/10.1038/35017641

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5, e9490. https://doi.org/10.1371/journal.pone.0009490

Ramirez, K.S., Leff, J.W., Barberán, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H., Fierer, N., 2014. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proc. R. Soc. B Biol. Sci. 281, 20141988. https://doi.org/10.1098/rspb.2014.1988

Rancurel, C., Legrand, L., Danchin, E., 2017. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the

Tree of Life. Genes 8, 248. https://doi.org/10.3390/genes8100248
Richard, P., Toivari, M.H., PenttilÃ¤, M., 2000. The role of xylulokinase in *Saccharomyces cerevisiae* xylulose catabolism. FEMS Microbiol. Lett. 190, 39–43. https://doi.org/10.1111/j.1574-6968.2000.tb09259.x
Rosso, M.-N., Favery, B., Piotte, C., Arthaud, L., De Boer, J.M., Hussey, R.S., Bakker, J., Baum, T.J., Abad, P., 1999. Isolation of a cDNA Encoding a β-1,4-endoglucanase in the Root-Knot Nematode *Meloidogyne incognita* and Expression Analysis During Plant Parasitism. Mol. Plant-Microbe Interactions® 12, 585–591. https://doi.org/10.1094/MPMI.1999.12.7.585
Scholl, E.H., Thorne, J.L., McCarter, J.P., Bird, D.M., 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. 4, R39. https://doi.org/10.1186/gb-2003-4-6-r39
Schwartz, J.A., Curtis, N.E., Pierce, S.K., 2014. FISH labeling reveals a horizontally transferred algal (Vaucheria litorea) nuclear gene on a sea slug (Elysia chlorotica) chromosome. Biol. Bull. 227, 300–312. https://doi.org/10.1086/BBLv227n3p300
Shin, N.R., Doucet, D., Pauchet, Y., 2022. Duplication of Horizontally Acquired GH5_2 Enzymes Played a Central Role in the Evolution of Longhorned Beetles. Mol. Biol. Evol. 39, msac128. https://doi.org/10.1093/molbev/msac128
Singh, S., Singh, B., Singh, A.P., 2015. Nematodes: A Threat to Sustainability of Agriculture. Procedia Environ. Sci. 29, 215–216. https://doi.org/10.1016/j.proenv.2015.07.270
Smant, G., Stokkermans, J.P.W.G., Yan, Y., de Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L., Helder, J., Schots, A., Bakker, J., 1998. Endogenous cellulases in animals: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc. Natl. Acad. Sci. 95, 4906–4911. https://doi.org/10.1073/pnas.95.9.4906
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding: NEXT-GENERATION DNA METABARCODING. Mol. Ecol. 21, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x
van Megen, H., van den Elsen, S., Holterman, M., Karssen, G., Mooyman, P., Bongers, T., Holovachov, O., Bakker, J., Helder, J., 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology 11, 927–950. https://doi.org/10.1163/156854109X456862
Watson, M., 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. https://doi.org/10.7488/ds/2296
Whitcomb, J.M., 1992. RETROVIRAL REVERSE TRANSCRIPTION AND INTEGRATION: Progress and Problems. Annu. Rev. Celt Bioi. 8, 275–306.

Acknowledgement

We thank MYCOPHYTO SAS and the Plant Health and environment department of the INRAe for supporting this project. We want to thank all members of the bioinformatics platform of the Institute Sophia Agrobiotech, Sophia Antipolis, France, for their help and support. We are also thankful for their advice on THG analyses Laura EME, Université Paris-Sud, Orsay, France and Samuel Mondy from INRAE, Dijon, France. We thank everyone who worked to generate the data publicly available on the IMG/M platform. Finally, we thank the Mesocentre OCA for providing powerful computing resources.

Code availability

Project name: HGT_in_PPN
Project home page: https://github.com/CaroleBelliardo/HGT_PPN_2022.git
Operating system(s): Platform independent
Programming language: Python3
Other requirements: Python3.8 or higher
License: GNU General Public License v3.0

Supplementary data

Table 1 | PPN data collection

Abbrev.	Species	BUSCO%	Proteins	Taxo	Hclade	Bclade	Lifestyle	Reference
Bxylo	*Bursaphelenchus xylophilus*	93.7	17704	Parasitaphelenchidae	10D	IV	Plant parasite	Kikuchi et all (2011)
Ddest	*Ditylenchus destructor*	87.1	13938	Tylenchida	12	IV	Plant parasite	Zheng et al (2016)
Ddips	*Ditylenchus dipsaci*	84.1	28064	Tylenchida	12	IV	Plant parasite	Mimee et al (2019)
Gpalli	*Globodera pallida*	72.6	16403	Tylenchida	12	IV	Plant parasite	Cotton et al (2014)
Gpaln	*Globodera pallida Newton*	96	16914	Tylenchida	12	IV	Plant parasite	Eves-van den Akker, in prep
Grosto	*Globodera rostochiensis*	93.4	14309	Tylenchida	12	IV	Plant parasite	Eves-van den Akker et al (2016)
Hglyc_1	*Heterodera glycines*	92.1	32270	Tylenchida	12	IV	Plant parasite	Masonbrink et al. (2019)
Hscha	*Heterodera schachtii*	92.5	32624	Tylenchida	12	IV	Plant parasite	Eves-van den Akker, in prep
Mare1	*Meloidogyne arenaria*	96.7	101269	Tylenchida	12	IV	Plant parasite	Blanc-Mathieu et al. (2017)
Ment3	*Meloidogyne incognita N.*	94.7	59773	Tylenchida	12	IV	Plant parasite	Koutsovoulos et al. (2020)
Mflor	*Meloidogyne floridensis*	70.3	21038	Tylenchida	12	IV	Plant parasite	Szitenberg et al. (2017)
Mgrami	*Meloidogyne graminicola*	88.1	10895	Tylenchida	12	IV	Plant parasite	Somvanshi et al (2018)
Mhapl	*Meloidogyne hapla*	90.8	14419	Tylenchida	12	IV	Plant parasite	Opperman et al (2008)
Minc3	*Meloidogyne incognita*	95.7	43718	Tylenchida	12	IV	Plant parasite	Blanc-Mathieu et al. (2017)
Mjav1	*Meloidogyne javanica*	95.3	97208	Tylenchida	12	IV	Plant parasite	Blanc-Mathieu et al. (2017)
Ppene	*Pratylenchus penetrans*	92.1	21950	Tylenchida	12	IV	Plant parasite	Paulo Vieira
Rreni	*Rotylenchulus reniformis*	82.5	41529	Tylenchida	12	IV	Plant parasite	Showmaker et al. (2019)
Rsimi	*Radopholus similis*	80.9	14817	Tylenchida	12	IV	Plant parasite	Wram et al. (2019)

Table 2 | Information about PPN proteomes and putative HGT classification

	Proteome						HGT detection		HGT Validation				
	nb. proteins		Protein length										
	Proteome	**Interpro**	**min**	**max**	**median**	**mean**	**AI | AHS > 0**	**% proteins**	**HGT**	**COMPLEX**	**likely_conta**	**pbAnnotDB**	**NO**
Bursaphelenchus_xylophilus	17704	16058	8	7699	293	375	1114	6,29	96	38	34	140	308
Ditylenchus_destructor	13938	12837	2	6485	322	430	927	6,65	176	48	4	84	312
Ditylenchus_dipsaci	28064	23331	17	5327	216	301	1965	7,00	412	107	32	265	816
Globodera_pallida	16403	14411	3	6633	250	359	1520	9,27	270	31	43	391	735
Globodera_pallida_Newton	16914	15170	50	11318	320	433	1389	8,21	270	33	1	298	602
Globodera_rostochiensis	14309	13192	2	6395	310	422	1019	7,12	237	20	5	124	386
Heterodera_glycines	32270	28337	13	5731	284	400	1391	4,31	324	35	1	216	576
Heterodera_schachtii	32624	27387	67	9221	279	370	2236	6,85	474	49	0	244	767
Meloidogyne_arenaria	101269	78670	49	6544	174	269	7510	7,42	1090	181	8	1225	2504
Meloidogyne_incognita_N.	59773	12541	39	5904	251	344	1393	6,62	853	106	5	1321	2285
Meloidogyne_graminicola	10895	17304	99	4820	332	432	912	8,37	208	34	57	213	512
Meloidogyne_hapla	14419	10096	8	5836	250	348	843	5,85	184	19	139	144	486
Meloidogyne_floridensis	21038	52863	1	3265	192	259	6036	10,1	134	28	3	102	267
Meloidogyne_incognita	43718	38969	49	6252	239	331	3709	8,48	574	98	2	608	1282
Meloidogyne_javanica	97208	75606	49	6150	169,5	256	6839	7,04	979	146	10	1083	2218
Pratylenchus_penetrans	21950	19535	62	6413	268	362	727	3,31	331	23	10	127	491
Rotylenchulus_reniformis	41529	35588	66	7186	191	303	2649	6,38	648	86	12	234	980
Radopholus_similis	14817	13814	29	6096	343	457	863	5,82	252	31	11	71	365

Figure 1 | Percentage of HGT per protein-coding genes

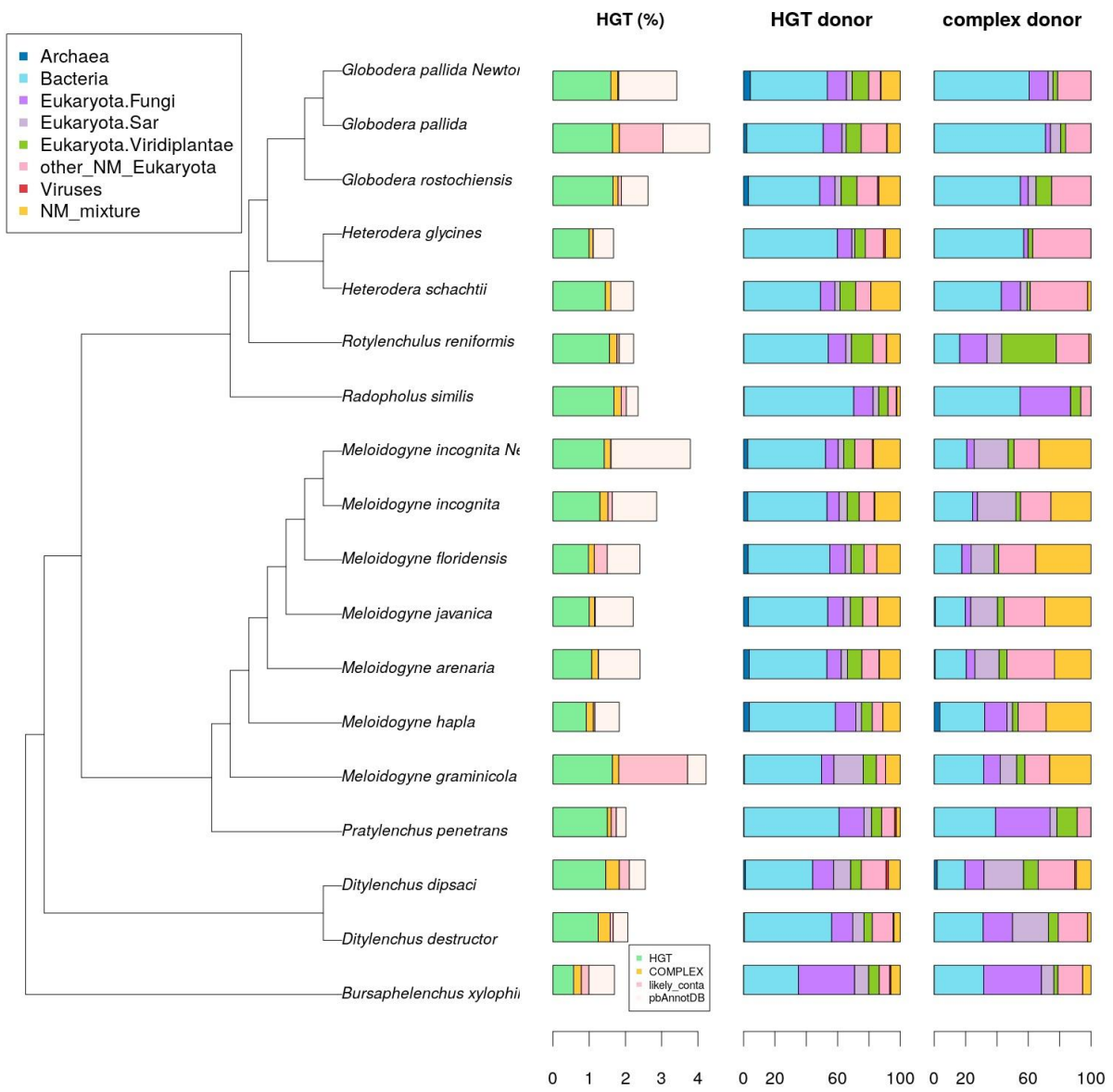


Figure 1 The phylogenetic tree of studied species displaying (a) the percentage of the predicted proteins classified as HGT in manually curated AvP results and (b) the kingdoms of potential donors of candidate HGT events, and (c) complex HGT events. The origins of donors were based on the sister branch taxa content. If no taxa account for 80% of all proteins in the sister branch, the donor is labelled as a non-metazoan mixture ('NM_mixture'). See material and method putative HGT classification rules.

Figure 2 | GH53 Phylogenetic tree

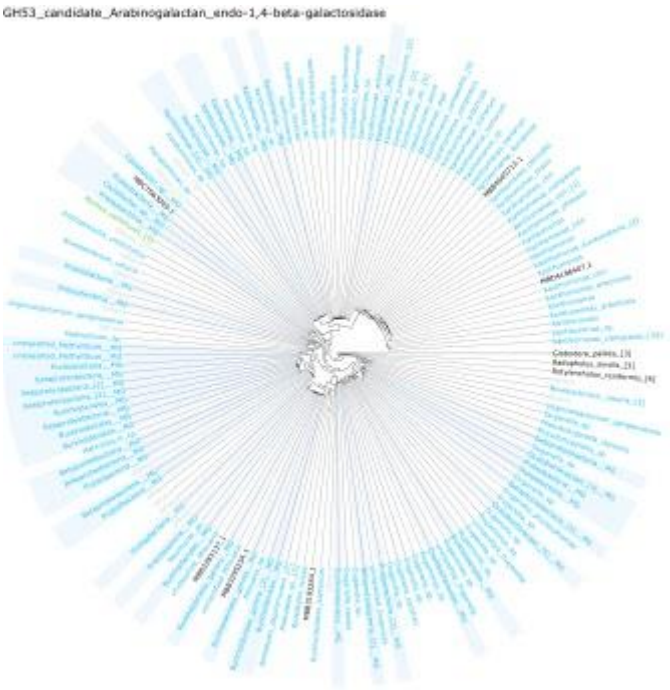


Figure 3 | Taxonomic rank of HGT events

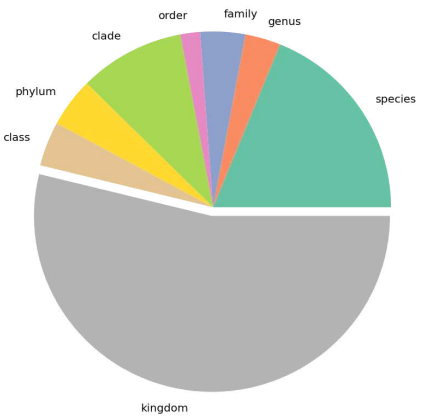


Figure 3 Taxonomic rank of HGT events putative donor LCA

Figure 4 | Number of intron per putative HGT

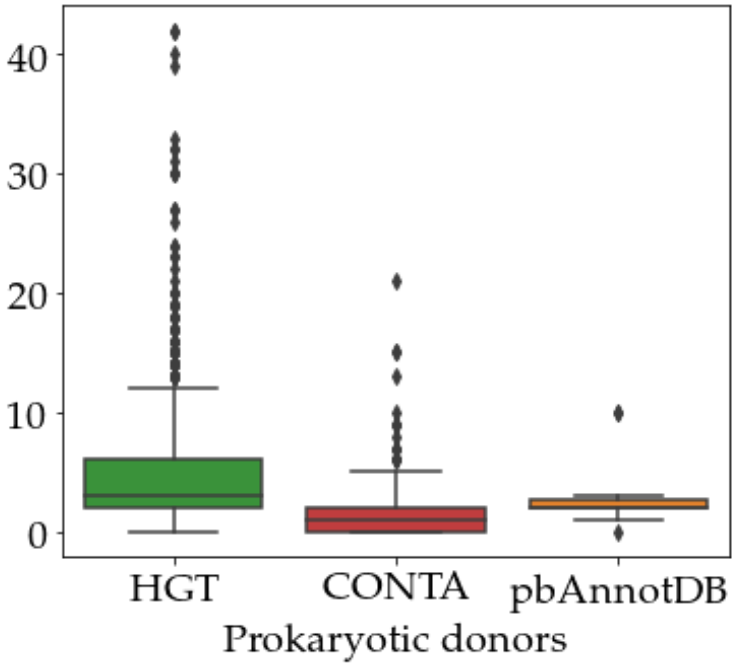


Figure 4 Number of intron per putative HGT of PNN according to whether they were phylogenetically validated (green), classified contamination (red) or related to the incorrect annotation in the reference database

3. *Conclusion et perspectives*

Après une curation manuelle des THG, nous avons déterminé qu'en moyenne ~1,3% des gènes codant pour des protéines avaient probablement été acquis par THG d'origine non animale chez les espèces de NPP du groupe *Tylenchoidea* et 0.5% chez le nématode parasite du pin. Les analyses phylogénétiques confirment qu' une majorité d'événements d'acquisition sont ancestraux aux groupes des *Tylenchoidea*. En outre, l'analyse des arbres phylogénétiques révèle que les protéines provenant d'échantillons environnementaux sont généralement plus étroitement apparentés aux gènes acquis par transferts horizontaux chez les NPP que les séquences provenant de bibliothèques généralistes. Les taxons les plus représentés parmi les branches soeurs des événements de THG et donc les possibles donneurs étaient les *Proteobacteria*, les *Actinobacteria*, les *Burkholderiaceae* et les *Rhizobiales* (4 clades de Bactéries), suivis des *Dikarya* (c'est-à-dire des espèces fongiques possédant 2 noyaux). Ces résultats confirment que les donneurs les plus probables étaient des micro-organismes du sol qui occupaient la même niche écologique que les ancêtres des NPP (Arima et Beppu, 1964 ; Fierer, 2017). Globalement, notre analyse a non seulement permis de valider les cas de THG déjà décrits dans la littérature mais de se rapprocher de l'identité des donneurs en enrichissant les branches soeurs par des données provenant de métagénomes de sol. Par ailleurs, nous avons également pu découvrir de nouveaux cas de THG ayant le même niveau de confiance et de soutien phylogénétique que ceux précédemment décrits. Ainsi, ce travail jette un éclairage sans précédent sur le rôle des THG dans l'histoire évolutive des nématodes parasites de plantes et de leur contribution à la composition de leur génome.

Discussion et perspectives

L'ensemble de ce travail a permis de préciser l'histoire évolutive des gènes acquis par THG chez les nématodes parasites de plantes et mettre en évidence l'étendue de ce phénomène dans les génomes de ces organismes.

L'intégration de données métagénomiques à notre librairie de séquences a permis de nettement enrichir la liste des protéines homologues aux gènes acquis par transferts horizontaux dans les génomes de NPP, et ainsi de compléter la liste des potentiels donneurs de ces gènes. En effet, dans de nombreux cas, les gènes de NPP sont plus proches d'un point de vue évolutif de protéines issues des métagénomes que du reste du vivant ce qui conforte l'hypothèse qu'une grande partie de ces gènes provient de micro-organismes telluriques comme suggéré dans les études précédentes (Danchin et al., 2010; Haegeman et al., 2011).

L'exhaustivité de notre librairie de séquences et les efforts réalisés pour améliorer la qualité des protéines et des assignations taxonomiques permettent d'assurer une grande fiabilité d'identification des protéines homologues. Cependant, l'origine de ces THG reste assez nébuleuse, avec de nombreux événements où les identifications les plus précises qui ont pu être définies sont à des rangs taxonomiques très ancestraux tels que Bacteria ou Fungi.

Une grande majorité des THG proviennent d'acquisitions anciennes, et en particulier ancestrales au clade *Tylenchidae*. Considérant la dynamique évolutive des génomes microbiens (Gibson et Eyre-Walker, 2019), les génomes ancestraux dont sont issus ces gènes étaient probablement assez différents de ceux des micro-organismes qui peuplent les sols d'aujourd'hui. Ainsi, une perspective intéressante qu'offre le fruit de ce travail serait la reconstruction des séquences protéiques ancestrales transférées originellement aux nématodes. Les méthodes de reconstruction des états ancestraux permettent, à partir des phylogénies et alignements multiples, la résurrection *in silico* de séquences provenant d'organismes aujourd'hui éteints. Ainsi, les séquences contemporaines des potentiels descendants peuvent être utilisées pour inférer les états des différents acides aminés qui constituaient la séquence protéique ancestrale (Gumulya et al., 2018). En enrichissant la liste des protéines homologues identifiables, les données métagénomiques devraient améliorer la résolution des prédictions. Après reconstruction, il est généralement possible d'envisager différentes analyses *in silico* pour prédire les caractéristiques fonctionnelles ancestrales de la protéine. Mieux encore, une synthèse du gène correspondant et son expression en bactérie ou levure et la purification du produit de ce gène permettraient de réaliser une étude fonctionnelle de la protéine ancestrale pour évaluer quelle a été la fonction d'origine de ces gènes et comment cette séquence aurait évolué chez les descendants des donneurs. Ce type d'analyse pourra aussi conforter l'hypothèse de l'acquisition par transferts horizontaux. En effet, deux types de reconstructions sont possibles : une reconstruction à partir des protéines non-métazoaires pour reconstruire la séquence microbienne, mais il est aussi possible d'envisager la reconstruction de la séquence ancestrale à partir des séquences de nématodes. La confrontation des deux modèles de séquence pourra mettre à l'épreuve nos conjectures concernant le processus de transfert horizontal.

D'autre part, on retrouve quelques dizaines de THG espèces spécifiques pour chacun des NPP étudiés. Des analyses complémentaires sont nécessaires pour valider l'insertion de ces gènes dans les génomes hôtes. Ces différentes validations devront se faire en plusieurs étapes. D'abord bio-informatiquement, par la recherche de ces gènes dans des données populationnelles ou à défaut dans d'autres versions de génomes. Si ces gènes sont retrouvés, cela pourra conforter l'hypothèse qu'ils font partie intégrante du génome de NPP. L'intégration pourra aussi être confirmée en laboratoire par une analyse PCR grâce au design d'amorces permettant d'amplifier spécifiquement cette région du génome. Il serait pertinent aussi de confirmer l'expression de ces gènes grâce à des données transcriptomiques. Si ces gènes sont bien présents dans les génomes de nématodes, ils représentent des cas extrêmement intéressants de THG car il s'agit probablement de transferts récents. Un tel cas représente l'opportunité sans précédent d'observer le devenir de gènes en provenance de micro-organismes juste après insertion dans un génome animal, et d'observer la possible domestication de ces gènes à l'œuvre. Aussi, l'analyse de l'environnement génomique de THG récents pourrait permettre de mieux comprendre le processus d'intégration et le contexte moléculaire propice à ce type d'événements dans les génomes de NPP. L'amélioration des assemblages de génomes de NPP permet également d'envisager une étude de la distribution des THG le long des chromosomes et de déterminer s'il existe certains hotspots d'insertion. Si c'est le cas, il pourrait être intéressant de déterminer si les régions en question possèdent certaines signatures caractéristiques (richesse en transposons, contenu en GC, etc.).

Enfin, ce travail constitue une analyse bio-informatique massive visant à initier de nouvelles découvertes mais ce type d'approche *in-silico* doit être validé *in vivo*. Avant d'aller plus loin, il est possible d'envisager certaines analyses bio-informatiques complémentaires. Tout d'abord, l'utilisation de l'outil Fasttree a permis de reconstruire rapidement des dizaines de milliers d'arbres phylogénétiques

et la détection comportant les potentiels THG. Cependant, d'autres outils tels que IQTREE (Minh et al., 2020) permettent de réaliser des reconstructions phylogénétiques au maximum de vraisemblance plus fiables mais beaucoup plus coûteuses en temps de calcul. Dans une telle analyse à grande échelle et sur des données aussi massives, les benchmarks réalisés entre ces deux outils donnaient des résultats assez similaires et ne justifiaient pas le temps de calcul bien supérieur. Maintenant que la liste d'arbres soutenant des événements de THG est restreinte, il serait intéressant de procéder aux reconstructions phylogénétiques avec IQTREE afin d'obtenir des topologies et valeurs de support des arbres encore plus solides qui permettraient d'affiner l'interprétation. D'autres part, l'outil AvP propose aussi d'effectuer de manière automatisée des analyses de topologie alternatives. L'idée est de comparer la vraisemblance des arbres soutenant l'hypothèse de THG à celles d'arbres alternatifs contraints ne soutenant pas cette hypothèse. Ce type d'analyse nécessiterait énormément de temps de calcul supplémentaire mais cela améliorerait encore le degré de confiance de nos résultats. Enfin, de nombreux cas classées à ce jour comme potentielles contaminations bénéficieraient de l'analyse de données transcriptomiques. En effet, si des données RNA-seq obtenues à partir d'ARN messagers sélectionnés sur la présence d'une queue poly-A s'alignent sur les génomes aux positions ces THG, l'hypothèse d'une contamination bactérienne devient improbable. Hormis les potentielles contaminations, les données transcriptomiques pourraient également nous renseigner sur le patron d'expression des gènes acquis par transferts horizontaux et ainsi nous en apprendre plus sur leurs potentielles fonctions.

Pour aller plus loin, concernant la validation des nouveaux cas de THG détectés, des analyses en laboratoire seront nécessaires. Des analyses PCR utilisant des amorces chevauchant une région génomique correspondant à un THG mais aussi à un gène de nématode permettrait de confirmer l'insertion des nouveaux THG identifiés. Cependant les données longues lectures provenant des génomes les plus

récents permettraient également de renforcer l'hypothèse de l'insertion dans le génome hôte si certaines d'entre elles couvrent à la fois le THG et un gène de l'hôte. Aussi, l'isolement de ce gène et l'expression hétérologue dans un vecteur d'expression pourraient permettre d'isoler la protéine puis de caractériser les fonctions biochimiques. L'analyse bio-informatique d'identification de peptide signal de sécrétion et des expériences d'immuno-localisation pourraient permettre de confirmer la sécrétion de la protéine et de la localiser chez la plante. Enfin, des expériences d'invalidation du gène (Knock-out) n'étant pas possibles aujourd'hui chez les NPP, l'interférence ARN (ou RNAi) pourrait permettre de préciser la fonction biologique de ces gènes.

Concernant les perspectives plus larges de ce travail, les données générées dans le cadre de ce projet pourront être utilisées par la communauté scientifique dans le cadre de différents travaux de recherche. La librairie de séquences de sol utilisée pour la détection des THG chez les NPP, est une ressource compacte et fiable de la diversité protéique microbienne. Il est prévu que cette ressource soit utilisée dans le cadre de travaux d'autres équipes de recherche portant sur le flux de gènes chez d'autres organismes telluriques, tels que les oomycètes phytopathogènes. D'autre part, la sous-partie de cet ensemble de données que représentent les protéines eucaryotes a déjà été publiée et rendue publique sur Recherche Data Gouv. Ainsi, elle pourra être utilisée plus largement par la communauté scientifique, dans le cadre de travaux portant sur les génomes des micro-organismes eucaryotes. Les centaines de milliers de protéines orphelines identifiées pourront être étudiées, plus en détail, pour tenter de prédire leurs caractéristiques biochimiques et potentielles fonctions biologiques en utilisant des outils tels que Alpha Fold (Jumper et al., 2021). Cependant, les résultats de ce type d'approche basée sur des méthodes d'apprentissage risquent d'être limités en raison de l'absence de similarité avec les séquences connues. Pour aller plus loin et mieux comprendre l'impact biologique de ces protéines dans les communautés microbiennes eucaryotes, la caractérisation en

laboratoire sera nécessaire et permettra peut-être de découvrir de nouvelles fonctions clés des écosystèmes terrestres.

À plus court terme, le travail de re-prédiction des protéines eucaryotes a contribué à la détection de certains THG d'origine eucaryote comme nous l'avons vu dans le deuxième chapitre. De nombreuses protéines provenant de ce travail de prédiction sont présentes dans les arbres phylogénétiques. On a d'une part des protéines provenant de micro-organismes eucaryotes qui enrichissent l'information concernant l'origine de ces gènes. D'autre part, de nombreuses protéines prédites dans le cadre de ce travail sont des protéines provenant d'organismes telluriques tels que des arthropodes et de nombreux nématodes. La présence de ce type de protéines enrichit les informations concernant le moment d'acquisition mais aussi l'histoire évolutive après insertion. Une ouverture possible de ces travaux consisterait à appliquer les mêmes méthodes à des métagénomes d'autres environnement d'intérêt (microbiote intestinal, milieu marin) pour lesquels les ressources abondent mais souffrent des mêmes biais de prédiction en faveur des procaryotes que les métagénomes de sol. Des développements méthodologiques pourraient même être envisagés pour, sur la base de ces travaux, améliorer la prédiction de gènes d'origine eucaryote directement au moment de la primo-analyse de métagénomes.

Aussi, comme expliqué dans le premier chapitre, le travail de re-prédiction des protéines eucaryotes a pu directement être transmis à Mycophyto pour alimenter les travaux de Recherche et Développement de cette société. Mycophyto commercialise des biostimulants agricoles composés principalement de CMA. Les solutions mycorhiziennes actuellement commercialisées par la concurrence sont des solutions mono-espèce. Ce type de produit est une solution non optimale car comme démontré par Crossay et collaborateurs, les communautés mycorhiziennes qui présentent une diversité d'espèces procurent un bénéfice accru sur la santé de la plante hôte (Crossay et al., 2019). Les solutions actuellement proposées par

Mycophyto reposent sur l'amplification de l'ensemble des CMA indigènes à partir de prélèvements du sol à enrichir, et sont ainsi constituées d'un ensemble d'espèces. Cette méthode permet d'optimiser la synergie plante-CMA grâce à l'ensemensement de solutions présentant une plus grande biodiversité. Cependant ce mode de production est plus lent en termes de production que la confection de solutions monospécifiques et plusieurs mois sont souvent nécessaires avant la mise en œuvre. Pour réduire le temps de production, Mycophyto souhaite développer un algorithme prédisant la composition mycorhizienne optimale en fonction des paramètres physico-chimiques de l'environnement. Théoriquement, si une espèce est présente dans plusieurs zones géographiques différentes présentant les mêmes caractéristiques, on peut faire l'hypothèse que cette espèce est particulièrement adaptée à ce milieu et peut développer une relation optimale avec la plante. En partant de cette hypothèse, ce type d'outil permettrait d'optimiser la production de mélanges d'espèces. Pour celà, la première étape était la constitution d'une base de données intégrant des informations pédologiques et climatiques mais aussi la répartition géographique des CMA. Dans le cadre d'une collaboration, l'équipe Acume de L'Institut national de recherche en sciences et technologies du numérique (INRIA) travaille actuellement sur le développement d'un algorithme de prédiction. Les milliers de métagénomes de sol géolocalisés, et les espèces de CMA que j'ai pu identifier dans ces sols constituent une mine d'informations sur la biodiversité tellurique, ainsi qu' une opportunité d'enrichir cette base de données.

L'utilisation de données de type métagénomique nécessite de travailler à une échelle protéique pour des raisons d'efficacité et d'optimisation. Les volumes de données sont drastiquement réduits ce qui permet l'utilisation d'outils plus fiables. Des outils bio-informatiques ont été développés pour réaliser des classifications taxonomiques à l'échelle génomique sur les gros volumes de données que représentent les métagénomes. C'est le cas notamment de KRAKEN (Wood et al., 2019), utilisé dans le travail du chapitre 1. Cet outil basé sur comparaison de K-mers

est classé dans les comparatifs comme le plus précis et rapide de sa catégorie (Ye et al., 2019). Ainsi, les outils de type recherche de similarité par alignement des régions homologues, tels que DIAMOND (Buchfink et al., 2015), présentent les meilleurs résultats en termes de sensibilité et de spécificité. Finalement, la prédiction des protéines avec des outils appropriés aux eucaryotes et leur annotation taxonomique avec une approche LCA constituent l'approche la plus fiable et efficace pour exploiter des données de type métagénomique shotgun pour la détection de CMA. En effet, comme mentionné dans l'article publié dans le journal Scientific data ce travail a permis d'améliorer la détection mais aussi de préciser l'information taxonomique des protéines provenant de CMA. Cependant, plusieurs limites subsistent dans ce projet. Premièrement, bien que plusieurs dizaines de milliers de protéines de CMA aient été identifiées, ces données ne constituent qu'un petit jeu de données pour le développement d'outils de prédiction et ont nécessité l'utilisation d'algorithmes adaptés. L'une des raisons est probablement l'encapsulation de l'ADN dans des spores très résistantes rendant le processus d'extraction d'ADN particulièrement difficile (Janoušková et al., 2015). Ainsi, la non-identification de CMA dans des données de séquençage ne signifie pas une absence de CMA dans les échantillons. Il est difficile d'évaluer l'importance de ce phénomène car le taux de faux négatifs n'a pas pu être évalué à partir des données publiques. Certains articles de méthodologie métagénomique utilisent des « mock » communautés, des communautés microbiennes de synthèse dont la composition est connue et contrôlée, pour évaluer le taux de faux positifs ressortant des approches métagénomiques shotgun (Naylor et al., 2020; Nicholls et al., 2019). Cependant, aucun article ne portant sur des *Glomeromycètes* n'a été publié à ce jour, et les résultats d'autres études « mock » peuvent difficilement être extrapolés aux CMA pour les raisons précitées. La deuxième limite est que les données métagénomiques shotgun ne permettent pas d'effectuer d'analyses quantitatives ou d'utiliser les algorithmes exploitant ce type d'informations. Pour améliorer ces prédictions, il sera donc nécessaire d'intégrer

davantage de données dans le futur. Considérant les proportions de données identifiées, les analyses de type metabarcoding seraient plus adaptées pour ce type d'analyses.). En effet, notre étude a permis de drastiquement améliorer la diversité des CMA identifiés dans les données métagénomiques mais l'identification ne couvre qu'une dizaine d'espèces alors que la littérature évalue à plusieurs centaines la diversité des CMA. En enrichissant la quantité de données tout en élargissant la diversité des CMA identifiés, les données de type metabarcoding permettront d'exploiter au mieux les ressources existantes comme la base de données Maarjam (REF OPIK) qui référence les séquences génomiques spécifiques à la classe des Gloméromycètes (taxon monophylétique incluant toutes les espèces de CMA), et d'optimiser la prédiction des mélanges de champignons mycorhiziens adaptés aux différents paramètres pédologiques et climatiques.

Concernant les autres perspectives qu'offrent les données générées et les résultats de ce travail, durant cette thèse 1 200 arbres phylogénétiques correspondant à des THG potentiels détectés par similarité ont été classés manuellement afin de définir les règles topologiques permettant d'affiner la validation des THG. Cette classification manuelle des THG potentiels constitue une excellente ressource pour développer un outil de détection des THG dans les génomes de NPP basée sur des méthodes d'intelligence artificielle. Les observations réalisées durant cette thèse confortent l'hypothèse que le développement de telles méthodes pourraient donner de bon résultats. En effet, les gènes acquis par transferts horizontaux dans les génomes de NPP semblent avoir des structures particulières. Par exemple, après validation phylogénétique les THG ont été classés en trois classes : « THG », « contamination » et « erreur d'annotation dans la librairie de référence ». La comparaison du nombre d'introns dans les gènes de ces différentes classes montre que le nombre d'introns est significativement plus élevé dans les THG que dans les gènes classés comme contamination (ce qui soutient l'hypothèse que ces gènes ont été domestiqués par le génome hôte PPN). En outre, lors de cette analyse nous avons

observé un nombre d'introns significativement plus élevé dans les THG que dans les gènes de la classe « erreur d'annotation dans la librairie de référence ». Hors, les gènes de cette dernière classe représentent bien des gènes de nématodes. Cette observation contre-intuitive pourrait suggérer que les THG contiennent un plus grand nombre d'introns que les gènes acquis verticalement. Cependant, à ce stade il s'agit d'une hypothèse qui pourrait être validée par comparaison du nombre d'introns dans les THG par rapport à tous les gènes codants dans l'ensemble du génome hôte. Si une telle différence dans la composition des introns/exons est observée spécifiquement dans les gènes acquis horizontalement, cela pourrait fournir de nouvelles informations sur les processus évolutifs impliqués dans la domestication des séquences bactériennes dans les génomes eucaryotes et sur la façon dont les gènes bactériens s'expriment dans les génomes eucaryotes.

L'utilisation de données métagénomiques de sol a permis d'élargir la détection des gènes acquis par THG dans les génomes de nématodes parasites de plantes et d'enrichir l'information sur l'origine de ces gènes. Notre compréhension de leur origine et de leur histoire évolutive reste encore partielle mais l'apport de la métagénomique dans cette étude est soutenu par la présence de nombreux gènes provenant de métagénomes dans les phylogénies et en particulier dans les branches sœurs des événements de THG. Ainsi, nos résultats indiquent qu'il sera pertinent d'intégrer des données environnementales aux futurs travaux portant sur la question des THG chez d'autres organismes. Pour aller plus loin dans la compréhension de l'implication des THG dans l'évolution du parasitisme des plantes, les prochaines études pourraient intégrer des nématodes ectoparasites ou des insectes parasites des parties aériennes des plantes. Par exemple, il serait intéressant d'étudier des nématodes du clade 2 tel que *Xiphinema* index, clade éloigné des *Tylenchina* d'un point de vue évolutif. Aucun génome de référence n'était disponible au début de ces travaux, mais une première version de très bonne qualité à un niveau de résolution chromosomique devrait être publiée très prochainement et devrait faire l'objet d'une

analyse des THG car la seule étude réalisée jusqu'à aujourd'hui portrait sur des données transcriptomiques (Danchin et al., 2017). Danchin et collaborateurs avaient montré la présence de gènes acquis par transferts horizontaux chez deux espèces de NPP du clade 2 dont l'espèce *X. index*. Cependant, une étude à l'échelle génomique offrira une vision plus précise de ce phénomène et permettra de mieux différencier les potentielles contamination des cas plus probables de THG. Plus largement, on sait aussi que de nombreuses espèces d'insectes phytophages telles que la mouche blanche *B. tabacci* ou les coléoptères de la famille *Chrysomeloidea* ont acquis des gènes par transferts horizontaux (Kirsch et al., 2014; Lapadula et al., 2020; Pauchet et al., 2010). L'intégration de données de métagénomiques provenant du séquençage d'échantillon de phyllosphère (parties des plantes situées au-dessus du niveau du sol) permettrait d'approfondir nos connaissances sur THG chez ces espèces en élargissant la détection mais aussi certainement le catalogue des potentiels donneurs. De même, concernant particulièrement les endoparasites, il serait intéressant d'inclure des données provenant de l'endophyte (microbiote présent à l'intérieur des plantes).

Compléter le catalogue des donneurs permet de préciser l'origine de ces gènes et pourrait guider notre compréhension des mécanismes impliqués dans le processus de transfert de matériel génétique entre espèces. La métagénomique environnementale est un domaine assez récent et plusieurs challenges méthodologiques subsistent notamment au niveau de la reconstruction des génomes. L'émergence de nouvelles technologies permet progressivement de relever ces challenges méthodologiques. Des données métagénomiques de meilleure qualité pourraient préciser l'origine de ces gènes mais aussi les mécanismes biologiques impliqués dans ce phénomène de THG. Retrouver les génomes de descendants des donneurs de ces gènes permettrait de préciser si ces gènes proviennent d'événements indépendants (un gène par événement) ou commun (plusieurs gènes acquis d'un même donneur lors d'un même événement).

Notre identification des donneurs est souvent réduite à des rangs taxonomiques ancestraux et ce en raison de la méthode utilisée pour re-annoter taxonomiquement l'ensemble des données métagénomiques. Les données métagénomiques de sols utilisées dans le cadre de cette étude sont principalement des données provenant de technologies de séquençage à lectures courtes. Plus de la moitié de ces données génomiques ont une taille inférieure ou égale à 250 pb (Belliardo et al., 2022). La reconstruction des génomes à partir de données métagénomiques obtenues à partir de courtes séquences est très difficile. La difficulté de cette tâche est d'autant plus forte pour des données provenant d'échantillons aussi complexes que le sol. La liste des solutions pour l'assignation taxonomique de données aussi fragmentées est très restreinte et des solutions telles que proposées par GTDB-tk ne sont pas envisageables. Des méthodes ont été développées pour classifier spécifiquement les espèces microbiennes sauvages telles que l'initiative GTDB qui a réalisé une classification phylogénétique des espèces procaryotes (Parks et al., 2022) et propose l'outil GTDB-tk (Chaumeil et al., 2020) qui propose le placement des nouveaux génomes dans cet arbre. Ce type de procédure est très lourd d'un point de vue informatique et n'est pas possible avec des données utilisées ici. L'approche LCA offre l'avantage d'une grande fiabilité des annotations, et c'est ce que nous avons choisi de privilégier dans cette première étude pionnière. Cependant, l'information fournie par cette approche est souvent peu précise.

Le développement de méthodes de séquençage troisième génération permettant de séquencer de bien plus longs fragments d'ADN peut minimiser la difficulté de la tâche d'assemblage. Originellement, les données issues de ces nouvelles technologies présentaient des taux d'erreur difficilement acceptables proche des 10%. Cependant, de nombreuses solutions émergent pour corriger ce taux d'erreur. C'est le cas notamment du séquençage pacbio Sequel II HiFI qui propose des lectures de haute qualité grâce à un procédé de circularisation des molécules d'ADN permettant d'effectuer de multiples lectures appelées *sous-reads* (Wenger et al., 2019). La

compilation de ces *sous-reads* permet de corriger les erreurs de séquençage, et d'obtenir un taux d'erreur théorique inférieur à 1%. Au début de cette thèse, nous avons effectué dix prélèvements de sols sur des parcelles agricoles connues pour être infectées par des nématodes parasites de plantes. Ces échantillons ont été séquencés avec une technologie longue lecture en utilisant la technologie Sequel II HiFI. L'exploitation de ces données reste encore à un stade préliminaire mais les lectures obtenues ont des tailles moyennes de 5kb bien supérieure à la plupart des contigs assemblés disponibles dans les données publiques. Les résultats préliminaires de la recherche de similarité des THG détectés chez les NPP indiquent la présence de nombreuses séquences homologues dans ces nouveaux métagénomes de haute qualité. Des méthodes de placement phylogénétiques pourront être utilisées pour replacer ces séquences homologues dans un contexte évolutif et leur relation de parenté avec les NPP. L'amélioration de la qualité des assemblages permettra d'envisager des méthodes d'assignation taxonomiques plus précises et d'accéder au contenu génétique des descendants des potentiels donneurs.

L'amélioration de la qualité des données métagénomiques pourrait encore améliorer notre compréhension de l'origine de ces gènes, mais il est important de garder à l'esprit que ces gènes ont été transférés dans les génomes d'un nématode ancestral à partir des génomes de micro-organismes ancestraux. Considérant la dynamique évolutive des micro-organismes (Gibson et Eyre-Walker, 2019; Ochman et Davalos, 2006), les génomes ancestraux dont sont issus ces gènes étaient probablement assez différents de ceux des micro-organismes qui peuplent les sols d'aujourd'hui. Ainsi, une des perspectives les plus intéressantes qu'offre le fruit de ce travail est la reconstruction de la séquence protéique ancestrale transférée.

Glossaire

Abiotique : lié au milieu, indépendant des êtres vivants.

Biotique : relatif au vivant.

Clade : groupement de plusieurs embranchements de plantes ou d'animaux ayant une organisation et une origine commune.

Gènes orthologues : gènes homologues issus d'événement de spéciation, par opposition aux gènes paralogues.

Gènes paralogues : gènes homologues issus d'événement de duplication, par opposition aux gènes orthologues.

Métazoaire : clade taxonomique regroupant l'ensemble des espèces animales, c'est-à-dire d'organismes eucaryotes multicellulaires mobiles et hétérotrophes.

Monophylétique: groupe ou taxon qui comprend un ancêtre et tous ses descendants.

Nématode : vers ronds non segmentés, pouvant être libres ou parasites d'animaux et ou de végétaux.

Paraphylétique : groupe ou taxon qui comprend un ancêtre et une partie de ses descendants.

Polyphylétique : groupe ou taxon défini par une ressemblance qui n'a pas été héritée d'un ancêtre commun.

Séquençage : détermination de l'ordre linéaire des composants d'une macromolécule. Séquençage génomique est le séquençage de la molécule d'ADN.

Topologie d'arbre : forme de l'arbre, ordre de branchement des nœuds.

Taxon : entité conceptuelle qui est censée regrouper tous les organismes vivants possédant en commun certains caractères taxonomiques bien définis comme l'espèce, le genre, la famille, l'ordre, etc.

Pression de sélection : désigne un phénomène qui se traduit par une évolution des espèces vivantes soumises à certaines contraintes environnementales.

Bibliographie

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M.-C., Coutinho, P.M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T.R., Markov, G.V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T.J., Blaxter, M., Bleve-Zacheo, T., Davis, E.L., Ewbank, J.J., Favery, B., Grenier, E., Henrissat, B., Jones, J.T., Laudet, V., Maule, A.G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J., Wincker, P., 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol. 26, 909–915. https://doi.org/10.1038/nbt.1482

Agrios, G.N., 2005. Plant pathology, 5th ed. ed. Elsevier Academic Press, Amsterdam ; Boston. ISBN: 9780120445653.

Ahmed, M., Roberts, N.G., Adediran, F., Smythe, A.B., Kocot, K.M., Holovachov, O., 2022. Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. Front. Ecol. Evol. 9, 769565. https://doi.org/10.3389/fevo.2021.769565

Alačević, M., 1963. Interspecific Recombination in Streptomyces. Nature 197, 1323–1323. https://doi.org/10.1038/1971323a0

Albalat, R., Cañestro, C., 2016. Evolution by gene loss. Nature Reviews Genetics 17, 379–391. https://doi.org/10.1038/nrg.2016.39

Alguacil, M., Díaz, G., Torres, P., Rodríguez-Caballero, G., Roldán, A., 2018. Host identity and functional traits determine the community composition of the arbuscular mycorrhizal fungi in facultative epiphytic plant species (preprint). Microbiology. https://doi.org/10.1101/307991

Altieri, M.A., 1999. The ecological role of biodiversity in agroecosystems. Agriculture, Ecosystems & Environment 74, 19–31. https://doi.org/10.1016/S0167-8809(99)00028-6

Anderson, N.G., 1970. Evolutionary Significance of Virus Infection. Nature 227, 1346–1347. https://doi.org/10.1038/2271346a0

André, H.M., Noti, M.-I., Lebrun, P., 1994. The soil fauna: the other last biotic frontier. Biodivers. Conserv. 3, 45–56. https://doi.org/10.1007/BF00115332

Arnaud, F., Varela, M., Spencer, T.E., Palmarini, M., 2008. Endogenous retroviruses: Coevolution of endogenous Betaretroviruses of sheep and their host. Cell. Mol. Life Sci. 65, 3422–3432. https://doi.org/10.1007/s00018-008-8500-9

Attwood, G.T., Herrera, F., Weissenstein, L.A., White, B.A., 1996. An endo-β-1,4-glucanase gene (*celA*) from the rumen anaerobe *Ruminococcus albus* 8: cloning, sequencing, and transcriptional analysis. Can. J. Microbiol. 42, 267–278. https://doi.org/10.1139/m96-039

Baalsrud, H.T., Tørresen, O.K., Solbakken, M.H., Salzburger, W., Hanel, R., Jakobsen, K.S., Jentoft, S., 2018. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol. Biol. Evol. 35, 593–606. https://doi.org/10.1093/molbev/msx311

Bai, H., Lester, G.M.S., Petishnok, L.C., Dean, D.A., 2017. Cytoplasmic transport and nuclear import of plasmid DNA. Biosci. Rep. 37, BSR20160616. https://doi.org/10.1042/BSR20160616

Bakker, J., Gommers, F., Smant, G., Abad, P., Rosso, M.-N., Dautova, M., 2001. Single pass cDNA sequencing - a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode Meloidogyne incognita. Nematology 3, 129–139. https://doi.org/10.1163/156854101750236259

Baldwin, J.G., Nadler, S.A., Adams, B.J., 2004. EVOLUTION OF PLANT PARASITISM AMONG NEMATODES. Annu. Rev. Phytopathol. 42, 83–105. https://doi.org/10.1146/annurev.phyto.42.012204.130804

Balzergue, C., 2012. Régulation de la symbiose endomycorhizienne par le phosphate 345.

Bardgett, R.D., van der Putten, W.H., 2014. Belowground biodiversity and ecosystem functioning. Nature 515, 505–511. https://doi.org/10.1038/nature13855

Bauters, L., Haegeman, A., Kyndt, T., Gheysen, G., 2014. Analysis of the transcriptome of *Hirschmanniella oryzae* to explore potential survival strategies and host-nematode interactions: Transcriptome analysis of *Hirschmanniella oryzae*. Mol. Plant Pathol. 15, 352–363. https://doi.org/10.1111/mpp.12098

Bebber, D.P., Ramotowski, M.A.T., Gurr, S.J., 2013. Crop pests and pathogens move polewards in a warming world. Nat. Clim. Change 3, 985–988. https://doi.org/10.1038/nclimate1990

Begun, D.J., Lindfors, H.A., Kern, A.D., Jones, C.D., 2007. Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba* / *Drosophila erecta* Clade. Genetics 176, 1131–1137. https://doi.org/10.1534/genetics.106.069245

Begun, D.J., Lindfors, H.A., Thompson, M.E., Holloway, A.K., 2006. Recently Evolved Genes Identified From *Drosophila yakuba* and *D. erecta* Accessory Gland Expressed Sequence Tags. Genetics 172, 1675–1681. https://doi.org/10.1534/genetics.105.050336

Bell, C.A., Magkourilou, E., Urwin, P.E., Field, K.J., 2022. Disruption of carbon for nutrient exchange between potato and arbuscular mycorrhizal fungi enhanced cyst nematode fitness and host pest tolerance. New Phytol. 234, 269–279. https://doi.org/10.1111/nph.17958

Belliardo, C., Koutsovoulos, G.D., Rancurel, C., Clément, M., Lipuma, J., Bailly-Bechet, M., Danchin, E.G.J., 2022. Improvement of eukaryotic protein predictions from soil metagenomes. Sci. Data 9, 311. https://doi.org/10.1038/s41597-022-01420-4

Béra-Maillet, C., Arthaud, L., Abad, P., Rosso, M.-N., 2000. Biochemical characterization of MI-ENG1, a family 5 endoglucanase secreted by the root-knot nematode Meloidogyne incognita: Root-knot nematode cellulase characterization. Eur. J. Biochem. 267, 3255–3263. https://doi.org/10.1046/j.1432-1327.2000.01356.x

Bernard and Pasteur on Alcoholic Fermentation, 1879. Edinb. Med. J. 25, 264.

Bhattacharya, D., Pelletreau, K.N., Price, D.C., Sarver, K.E., Rumpho, M.E., 2013. Genome Analysis of Elysia chlorotica Egg DNA Provides No Evidence for Horizontal Gene Transfer into the Germ Line of This Kleptoplastic Mollusc. Mol. Biol. Evol. 30, 1843–1852. https://doi.org/10.1093/molbev/mst084

Bianchini, K., Tattersall, G.J., Sashaw, J., Porteus, C.S., Wright, P.A., 2012. Acid Water Interferes with Salamander–Green Algae Symbiosis during Early Embryonic Development. Physiol. Biochem. Zool. 85, 470–480. https://doi.org/10.1086/667407

Bitto, N.J., Chapman, R., Pidot, S., Costin, A., Lo, C., Choi, J., D'Cruze, T., Reynolds, E.C., Dashper, S.G., Turnbull, L., Whitchurch, C.B., Stinear, T.P., Stacey, K.J., Ferrero, R.L., 2017. Bacterial membrane vesicles transport their DNA cargo into host cells. Sci. Rep. 7, 7072. https://doi.org/10.1038/s41598-017-07288-4

Blaxter, M., Koutsovoulos, G., 2015. The evolution of parasitism in Nematoda. Parasitology 142 Suppl 1, S26-39. https://doi.org/10.1017/S0031182014000791

Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M., Vida, J.T., Thomas, W.K., 1998. A molecular evolutionary framework for the phylum Nematoda. Nature 392, 71–75. https://doi.org/10.1038/32160

Bonfante, P., Genre, A., 2008. Plants and arbuscular mycorrhizal fungi: an evolutionary-developmental perspective. Trends Plant Sci. 13, 492–498. https://doi.org/10.1016/j.tplants.2008.07.001

Bonkowski, M., 2004. Protozoa and plant growth: the microbial loop in soil revisited. New Phytol. 162, 617–631. https://doi.org/10.1111/j.1469-8137.2004.01066.x

Brochier-Armanet, C., Boussau, B., Gribaldo, S., Forterre, P., 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat. Rev. Microbiol. 6, 245–252. https://doi.org/10.1038/nrmicro1852

Brown, A.M.V., Wasala, S.K., Howe, D.K., Peetz, A.B., Zasada, I.A., Denver, D.R., 2018. Comparative Genomics of Wolbachia–Cardinium Dual Endosymbiosis in a Plant-Parasitic Nematode. Front. Microbiol. 9, 2482. https://doi.org/10.3389/fmicb.2018.02482

Campbell, N.A., Reece, J.B., Mathieu, R., 2004. Biologie, 2e éd. ed. ERPI, St-Laurent, Québec.

Campolino, M.L., de Paula Lana, U.G., Gomes, E.A., Coelho, A.M., de Sousa, S.M., 2022. Phosphate fertilization affects rhizosphere microbiome of maize and sorghum genotypes. Braz. J. Microbiol. Publ. Braz. Soc. Microbiol. 53, 1371–1383. https://doi.org/10.1007/s42770-022-00747-9

Candido, V., Campanelli, G., D'Addabbo, T., Castronuovo, D., Renco, M., Camele, I., 2013. Growth and yield promoting effect of artificial mycorrhization combined with different fertiliser rates on field-grown tomato. Ital. J. Agron. 8, 22. https://doi.org/10.4081/ija.2013.e22

Castagnone-Sereno, P., Mulet, K., Danchin, E.G.J., Koutsovoulos, G.D., Karaulic, M., Da Rocha, M., Bailly-Bechet, M., Pratx, L., Perfus-Barbeoch, L., Abad, P., 2019. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. Mol. Ecol. 28, 2559–2572. https://doi.org/10.1111/mec.15095

Catoni, M., Noris, E., Vaira, A.M., Jonesman, T., Matić, S., Soleimani, R., Behjatnia, S.A.A., Vinals, N., Paszkowski, J., Accotto, G.P., 2018. Virus-mediated export of chromosomal DNA in plants. Nat. Commun. 9, 5308. https://doi.org/10.1038/s41467-018-07775-w

Cayrol, J.C., 2013. Biological control of Meloidogyne by Arthrobotrys irregularis 6, 265–273.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, I.-M.A., Markowitz, V.M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., Varghese, N., Hadjithomas, M., Tennessen, K., Nielsen, T., Ivanova, N.N., Kyrpides, N.C., 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 45, D507–D516. https://doi.org/10.1093/nar/gkw929

Chen, L., Zheng, Y., Gao, C., Mi, X.-C., Ma, K.-P., Wubet, T., Guo, L.-D., 2017. Phylogenetic relatedness explains highly interconnected and nested symbiotic networks of woody

plants and arbuscular mycorrhizal fungi in a Chinese subtropical forest. Mol. Ecol. 26, 2563–2575. https://doi.org/10.1111/mec.14061

Chen, M., Arato, M., Borghi, L., Nouri, E., Reinhardt, D., 2018. Beneficial Services of Arbuscular Mycorrhizal Fungi – From Ecology to Application. Front. Plant Sci. 9, 1270. https://doi.org/10.3389/fpls.2018.01270

Chitwood, B.G., 1958. The designation of official names for higher taxa of invertebrates. Bull. Zool. Nomencl. 15B, 860–895. https://doi.org/10.5962/bhl.part.19410

Corradi, N., 2015. Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. Annu. Rev. Microbiol. 69, 167–183. https://doi.org/10.1146/annurev-micro-091014-104136

Cotton, J.A., Lilley, C.J., Jones, L.M., Kikuchi, T., Reid, A.J., Thorpe, P., Tsai, I.J., Beasley, H., Blok, V., Cock, P.J., Eves-van den Akker, S., Holroyd, N., Hunt, M., Mantelin, S., Naghra, H., Pain, A., Palomares-Rius, J.E., Zarowiecki, M., Berriman, M., Jones, J.T., Urwin, P.E., 2014. The genome and life-stage specific transcriptomes of Globodera pallida elucidate key aspects of plant parasitism by a cyst nematode. Genome Biol. 15, R43. https://doi.org/10.1186/gb-2014-15-3-r43

Craig, J.P., Bekal, S., Hudson, M., Domier, L., Niblack, T., Lambert, K.N., 2008. Analysis of a Horizontally Transferred Pathway Involved in Vitamin B6 Biosynthesis from the Soybean Cyst Nematode Heterodera glycines. Mol. Biol. Evol. 25, 2085–2098. https://doi.org/10.1093/molbev/msn141

Crossay, T., Majorel, C., Redecker, D., Gensous, S., Medevielle, V., Durrieu, G., Cavaloc, Y., Amir, H., 2019. Is a mixture of arbuscular mycorrhizal fungi better for plant growth than single-species inoculants? Mycorrhiza 29, 325–339. https://doi.org/10.1007/s00572-019-00898-y

Danchin, E., Perfus-Barbeoch, L., Rancurel, C., Thorpe, P., Da Rocha, M., Bajew, S., Neilson, R., (Guzeeva), E.S., Da Silva, C., Guy, J., Labadie, K., Esmenjaud, D., Helder, J., Jones, J., den Akker, S., 2017. The Transcriptomes of Xiphinema index and Longidorus elongatus Suggest Independent Acquisition of Some Plant Parasitism Genes by Horizontal Gene Transfer in Early-Branching Nematodes. Genes 8, 287. https://doi.org/10.3390/genes8100287

Danchin, E.G.J., Arguel, M.-J., Campan-Fournier, A., Perfus-Barbeoch, L., Magliano, M., Rosso, M.-N., Da Rocha, M., Da Silva, C., Nottet, N., Labadie, K., Guy, J., Artiguenave, F., Abad, P., 2013. Identification of Novel Target Genes for Safer and More Specific Control of Root-Knot Nematodes from a Pan-Genome Mining. PLoS Pathog. 9, e1003745. https://doi.org/10.1371/journal.ppat.1003745

Danchin, E.G.J., Guzeeva, E.A., Mantelin, S., Berepiki, A., Jones, J.T., 2016. Horizontal Gene Transfer from Bacteria Has Enabled the Plant-Parasitic Nematode *Globodera pallida* to Feed on Host-Derived Sucrose. Mol. Biol. Evol. 33, 1571–1579. https://doi.org/10.1093/molbev/msw041

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., Abad, P., 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. 201008486. https://doi.org/10.1073/pnas.1008486107

Dastogeer, K.M.G., Yasuda, M., Okazaki, S., 2022. Microbiome and pathobiome analyses reveal changes in community structure by foliar pathogen infection in rice. Front. Microbiol. 13, 949152. https://doi.org/10.3389/fmicb.2022.949152

Dautova, M., Rosso, M.-N., Abad, P., Gommers, F., Bakker, J., Smant, G., 2001. Single pass cDNA sequencing - a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode Meloidogyne incognita. Nematology 3, 129–139. https://doi.org/10.1163/156854101750236259

De Ley, P., 2006. A quick tour of nematode diversity and the backbone of nematode phylogeny. WormBook. https://doi.org/10.1895/wormbook.1.41.1

Deng, C., Cheng, C.-H.C., Ye, H., He, X., Chen, L., 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc. Natl. Acad. Sci. 107, 21593–21598. https://doi.org/10.1073/pnas.1007883107

Detrey, J., Cognard, V., Djian-Caporalino, C., Marteu, N., Doidy, J., Pourtau, N., Vriet, C., Maurousset, L., Bouchon, D., Clause, J., 2022. Growth and root-knot nematode infection of tomato are influenced by mycorrhizal fungi and earthworms in an intercropping cultivation system with leeks. Appl. Soil Ecol. 169, 104181. https://doi.org/10.1016/j.apsoil.2021.104181

Djian-Caporalino, C., Navarrete, M., Fazari, A., Baily-Bechet, M., Marteu, N., Dufils, A., Tchamitchian, M., Lefèvre, A., Pares, L., Mateille, T., Tavoillot, J., (‡) A.P., Sage-Palloix, A.-M., Védie, H., Goillon, C., Castagnone-Sereno, P., 2019. Conception et évaluation de systèmes de culture maraîchers méditerranéens innovants pour gérer les nématodes à galles. https://doi.org/10.25518/1780-4507.17725

Dong Wei, Y., Sik Lee, K., Zheng Gui, Z., Joo Yoon, H., Kim, I., Zheng Zhang, G., Guo, X., Dae Sohn, H., Rae Jin, B., 2006. Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, Apriona germari. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 145, 220–229. https://doi.org/10.1016/j.cbpb.2006.07.007

Douanne, N., Dong, G., Amin, A., Bernardo, L., Blanchette, M., Langlais, D., Olivier, M., Fernandez-Prada, C., 2022. Leishmania parasites exchange drug-resistance genes through extracellular vesicles. Cell Rep. 40, 111121. https://doi.org/10.1016/j.celrep.2022.111121

Downing, T., Imamura, H., Decuypere, S., Clark, T.G., Coombs, G.H., Cotton, J.A., Hilley, J.D., de Doncker, S., Maes, I., Mottram, J.C., Quail, M.A., Rijal, S., Sanders, M., Schönian, G., Stark, O., Sundar, S., Vanaerschot, M., Hertz-Fowler, C., Dujardin, J.-C., Berriman, M., 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug

resistance. Genome Res. 21, 2143–2156. https://doi.org/10.1101/gr.123430.111

Doyle, E.A., Lambert, K.N., 2002. Cloning and Characterization of an Esophageal-Gland-Specific Pectate Lyase from the Root-Knot Nematode *Meloidogyne javanica*. Mol. Plant-Microbe Interactions® 15, 549–556. https://doi.org/10.1094/MPMI.2002.15.6.549

Dropkin, V.H., 1963. Cellulase in Phytoparasitic Nematodes. Nematologica 9, 444–454. https://doi.org/10.1163/187529263X00980

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B.A., Rivero, F., Bankier, A.T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M.A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E.C., Chisholm, R.L., Gibbs, R., Loomis, W.F., Platzer, M., Kay, R.R., Williams, J., Dear, P.H., Noegel, A.A., Barrell, B., Kuspa, A., 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435, 43–57. https://doi.org/10.1038/nature03481

Elling, A.A., Mitreva, M., Gai, X., Martin, J., Recknor, J., Davis, E.L., Hussey, R.S., Nettleton, D., McCarter, J.P., Baum, T.J., 2009. Sequence mining and transcript profiling to explore cyst nematode parasitism. BMC Genomics 10, 58. https://doi.org/10.1186/1471-2164-10-58

Ellison, C.K., Dalia, T.N., Vidal Ceballos, A., Wang, J.C.-Y., Biais, N., Brun, Y.V., Dalia, A.B., 2018. Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. Nat. Microbiol. 3, 773–780. https://doi.org/10.1038/s41564-018-0174-y

Eme, L., Spang, A., Lombard, J., Stairs, C.W., Ettema, T.J.G., 2017. Archaea and the origin of eukaryotes. Nat. Rev. Microbiol. 15, 711–723. https://doi.org/10.1038/nrmicro.2017.133

Eom, A.-H., Hartnett, D.C., Wilson, G.W.T., 2000. Host plant species effects on arbuscular mycorrhizal fungal communities in tallgrass prairie. Oecologia 122, 435–444. https://doi.org/10.1007/s004420050050

Estrada, B., Aroca, R., Maathuis, F.J.M., Barea, J.M., Ruiz-Lozano, J.M., 2013. Arbuscular mycorrhizal fungi native from a Mediterranean saline area enhance maize tolerance to salinity through improved ion homeostasis. Plant Cell Environ. 36, 1771–1782.

https://doi.org/10.1111/pce.12082

Evangelisti, E., Govetto, B., Minet-Kebdani, N., Kuhn, M., Attard, A., Ponchet, M., Panabières, F., Gourgues, M., 2013. The *P hytophthora parasitica* RXLR effector Penetration-Specific Effector 1 favours *A rabidopsis thaliana* infection by interfering with auxin physiology. New Phytol. 199, 476–489. https://doi.org/10.1111/nph.12270

Fei, W., Liu, Y., 2022. Biotrophic Fungal Pathogens: a Critical Overview. Appl. Biochem. Biotechnol. https://doi.org/10.1007/s12010-022-04087-0

Ferraz, L.C.C.B., Brown, D.J.F., 2002. An introduction to nematodes: plant nematology ; a student's textbook, Pensoft series parasitologica. Pensoft Publishers, Sofia.

Fierer, N., 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579–590. https://doi.org/10.1038/nrmicro.2017.87

Filip, E., Skuza, L., 2021. Horizontal Gene Transfer Involving Chloroplasts. Int. J. Mol. Sci. 22, 4484. https://doi.org/10.3390/ijms22094484

Finlay, B.J., Esteban, G.F., 2001. Exploring Leeuwenhoek's legacy: the abundance and diversity of protozoa. Int. Microbiol. 4, 125–133. https://doi.org/10.1007/s10123-001-0027-y

Fitzpatrick, D.A., 2012. Horizontal gene transfer in fungi. FEMS Microbiol. Lett. 329, 1–8. https://doi.org/10.1111/j.1574-6968.2011.02465.x

Fleming, L.E., Bean, J.A., Rudolph, M., Hamilton, K., 1999. Mortality in a cohort of licensed pesticide applicators in Florida. Occup. Environ. Med. 56, 14–21. https://doi.org/10.1136/oem.56.1.14

Furuya, E.Y., Lowy, F.D., 2006. Antimicrobial-resistant bacteria in the community setting. Nat. Rev. Microbiol. 4, 36–45. https://doi.org/10.1038/nrmicro1325

Gao, B., Allen, R., Maier, T., McDermott, J.P., Davis, E.L., Baum, T.J., Hussey, R.S., 2002. Characterisation and developmental expression of a chitinase gene in Heterodera glycines. Int. J. Parasitol. 32, 1293–1300. https://doi.org/10.1016/S0020-7519(02)00110-8

Garcia, N., Grenier, E., Buisson, A., Folcher, L., 2022. Diversity of plant parasitic nematodes characterized from fields of the French national monitoring programme for the Columbia root-knot nematode. PLOS ONE 17, e0265070. https://doi.org/10.1371/journal.pone.0265070

Garzo, E., Rizzo, E., Fereres, A., Gomez, S.K., 2020. High levels of arbuscular mycorrhizal fungus colonization on *Medicago truncatula* reduces plant suitability as a host for pea aphids (*Acyrthosiphon pisum*). Insect Sci. 27, 99–112. https://doi.org/10.1111/1744-7917.12631

Georgelis, N., Nikolaidis, N., Cosgrove, D.J., 2015. Bacterial expansins and related proteins from the world of microbes. Appl. Microbiol. Biotechnol. 99, 3807–3823. https://doi.org/10.1007/s00253-015-6534-0

Gibson, B., Eyre-Walker, A., 2019. Investigating Evolutionary Rate Variation in Bacteria. J. Mol. Evol. 87, 317–326. https://doi.org/10.1007/s00239-019-09912-5

Gilbert, C., Belliardo, C., 2022. The diversity of endogenous viral elements in insects. Curr. Opin. Insect Sci. 49, 48–55. https://doi.org/10.1016/j.cois.2021.11.007

Glass, N.L., Schmoll, M., Cate, J.H.D., Coradetti, S., 2013. Plant Cell Wall Deconstruction by Ascomycete Fungi. Annu. Rev. Microbiol. 67, 477–498. https://doi.org/10.1146/annurev-micro-092611-150044

Gomes Carneiro, R.M.D., Cayrol, J.C., 1991. Relationship between inoculum density of the nematophagous fungus Paecilomyces lilacinus and control of Meloidogyne arenaria on tomato. Rev. Nématologie 14, 629–634.

Graham, L.A., Davies, P.L., 2021. Horizontal Gene Transfer in Vertebrates: A Fishy Tale. Trends Genet. 37, 501–503. https://doi.org/10.1016/j.tig.2021.02.006

Graham, L.A., Li, J., Davidson, W.S., Davies, P.L., 2012. Smelt was the likely beneficiary of an antifreeze gene laterally transferred between fishes. BMC Evol. Biol. 12, 190. https://doi.org/10.1186/1471-2148-12-190

Green, B.J., Li, W.-Y., Manhart, J.R., Fox, T.C., Summer, E.J., Kennedy, R.A., Pierce, S.K., Rumpho, M.E., 2000. Mollusc-Algal Chloroplast Endosymbiosis. Photosynthesis, Thylakoid Protein Maintenance, and Chloroplast Gene Expression Continue for Many Months in the Absence of the Algal Nucleus. Plant Physiol. 124, 331–342.

Griffith, F., 1928. The Significance of Pneumococcal Types. J. Hyg. (Lond.) 27, 113–159. https://doi.org/10.1017/S0022172400031879

Grynberg, P., Coiti Togawa, R., Dias de Freitas, L., Antonino, J.D., Rancurel, C., Mota do Carmo Costa, M., Grossi-de-Sa, M.F., Miller, R.N.G., Brasileiro, A.C.M., Messenberg Guimaraes, P., Danchin, E.G.J., 2020. Comparative Genomics Reveals Novel Target Genes towards Specific Control of Plant-Parasitic Nematodes. Genes 11, 1347. https://doi.org/10.3390/genes11111347

Gumulya, Y., Baek, J.-M., Wun, S.-J., Thomson, R.E.S., Harris, K.L., Hunter, D.J.B., Behrendorff, J.B.Y.H., Kulig, J., Zheng, S., Wu, X., Wu, B., Stok, J.E., De Voss, J.J., Schenk, G., Jurva, U., Andersson, S., Isin, E.M., Bodén, M., Guddat, L., Gillam, E.M.J., 2018. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. Nat. Catal. 1, 878–888. https://doi.org/10.1038/s41929-018-0159-5

Haegeman, A., Jones, J.T., Danchin, E.G.J., 2011. Horizontal gene transfer in nematodes: a catalyst for plant parasitism? Mol. Plant-Microbe Interact. MPMI 24, 879–887. https://doi.org/10.1094/MPMI-03-11-0055

Han, L., Xu, M., Kong, X., Liu, X., Wang, Q., Chen, G., Xu, K., Nie, J., 2022. Deciphering the diversity, composition, function, and network complexity of the soil microbial community after repeated exposure to a fungicide boscalid. Environ. Pollut. 312, 120060. https://doi.org/10.1016/j.envpol.2022.120060

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245-249. https://doi.org/10.1016/s1074-5521(98)90108-9

Hare, W., 1956. Resistance in pepper to Meloidogyne incognita acrita. Phytopathology 46, 98–104.

Hendy, H., Pochard, E., Dalmasso, A., Bongiovanni, M., 1985. Transmission héréditaire de la résistance aux nématodes Meloidogyne Chitwood (*Tylenchida*) portée par 2 lignées de Capsicum annuum L. : étude de descendances homozygotes issues d'androgenèse. Agron. EDP Sci. 93-100.

Henrissat, B., Bairoch, A., 1993. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. 293 (Pt 3), 781–788. https://doi.org/10.1042/bj2930781

Hestrin, R., Kan, M., Lafler, M., Wollard, J., Kimbrel, J.A., Ray, P., Blazewicz, S.J., Stuart, R., Craven, K., Firestone, M., Nuccio, E.E., Pett-Ridge, J., 2022. Plant-associated fungi support bacterial resilience following water limitation. ISME J. https://doi.org/10.1038/s41396-022-01308-6

Hijri, I., Sýkorová, Z., Oehl, F., Ineichen, K., Mäder, P., Wiemken, A., Redecker, D., 2006. Communities of arbuscular mycorrhizal fungi in arable soils are not necessarily low in diversity. Mol. Ecol. 15, 2277–2289. https://doi.org/10.1111/j.1365-294X.2006.02921.x

Holterman, M., Karegar, A., Mooijman, P., van Megen, H., van den Elsen, S., Vervoort, M.T.W., Quist, C.W., Karssen, G., Decraemer, W., Opperman, C.H., Bird, D.M., Kammenga, J., Goverse, A., Smant, G., Helder, J., 2017. Disparate gain and loss of parasitic abilities among nematode lineages. PLOS ONE 12, e0185445. https://doi.org/10.1371/journal.pone.0185445

Holterman, M., Karssen, G., van den Elsen, S., van Megen, H., Bakker, J., Helder, J., 2009. Small Subunit rDNA-Based Phylogeny of the *Tylenchida* Sheds Light on Relationships Among Some High-Impact Plant-Parasitic Nematodes and the Evolution of Plant Feeding. Phytopathology® 99, 227–235. https://doi.org/10.1094/PHYTO-99-3-0227

Holterman, M., van der Wurff, A., van den Elsen, S., van Megen, H., Bongers, T., Holovachov, O., Bakker, J., Helder, J., 2006. Phylum-Wide Analysis of SSU rDNA Reveals Deep Phylogenetic Relationships among Nematodes and Accelerated Evolution toward Crown Clades. Mol. Biol. Evol. 23, 1792–1800. https://doi.org/10.1093/molbev/msl044

Hunter, M.C., Smith, R.G., Schipanski, M.E., Atwood, L.W., Mortensen, D.A., 2017. Agriculture in 2050: Recalibrating Targets for Sustainable Intensification. BioScience 67, 386–391. https://doi.org/10.1093/biosci/bix010

Husnik, F., McCutcheon, J.P., 2018. Functional horizontal gene transfer from bacteria to eukaryotes. Nat. Rev. Microbiol. 16, 67–79. https://doi.org/10.1038/nrmicro.2017.137

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Janoušková, M., Püschel, D., Hujslová, M., Slavíková, R., Jansa, J., 2015. Quantification of arbuscular mycorrhizal fungal DNA in roots: how important is material preservation? Mycorrhiza 25, 205–214. https://doi.org/10.1007/s00572-014-0602-7

Jansson, J.K., Hofmockel, K.S., 2020. Soil microbiomes and climate change. Nat. Rev. Microbiol. 18, 35–46. https://doi.org/10.1038/s41579-019-0265-7

Jardillier, L., Zubkov, M.V., Pearman, J., Scanlan, D.J., 2010. Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. ISME J. 4, 1180–1192. https://doi.org/10.1038/ismej.2010.36

Jaubert, S., Laffaire, J.-B., Abad, P., Rosso, M.-N., 2002. A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne incognita* [1]. FEBS Lett. 522, 109–112. https://doi.org/10.1016/S0014-5793(02)02906-X

Jones, J.T., Furlanetto, C., Bakker, E., Banks, B., Blok, V., Chen, Q., Phillips, M., Prior, A., 2003. Characterization of a chorismate mutase from the potato cyst nematode *Globodera pallida*: *G. pallida* chorismate mutase. Mol. Plant Pathol. 4, 43–50. https://doi.org/10.1046/j.1364-3703.2003.00140.x

Jones, J.T., Haegeman, A., Danchin, E.G.J., Gaur, H.S., Helder, J., Jones, M.G.K., Kikuchi, T., Manzanilla-López, R., Palomares-Rius, J.E., Wesemael, W.M.L., Perry, R.N., 2013. Top 10 plant-parasitic nematodes in molecular plant pathology: Top 10 plant-parasitic nematodes. Mol. Plant Pathol. 14, 946–961. https://doi.org/10.1111/mpp.12057

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., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kambayashi, C., Kakehashi, R., Sato, Y., Mizuno, H., Tanabe, H., Rakotoarison, A., Künzel, S., Furuno, N., Ohshima, K., Kumazawa, Y., Nagy, Z.T., Mori, A., Allison, A., Donnellan, S.C., Ota, H., Hoso, M., Yanagida, T., Sato, H., Vences, M., Kurabayashi, A., 2022. Geography-Dependent Horizontal Gene Transfer from Vertebrate Predators to Their Prey. Mol. Biol. Evol. 39, msac052. https://doi.org/10.1093/molbev/msac052

Kasuya, M., 1964. TRANSFER OF DRUG RESISTANCE BETWEEN ENTERIC BACTERIA INDUCED IN THE MOUSE INTESTINE. J. Bacteriol. 88, 322–328.

https://doi.org/10.1128/jb.88.2.322-328.1964

Kikuchi, T., Jones, J.T., Aikawa, T., Kosaka, H., Ogura, N., 2004. A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus*. FEBS Lett. 572, 201–205. https://doi.org/10.1016/j.febslet.2004.07.039

King, A.M., Aaron, C.K., 2015. Organophosphate and Carbamate Poisoning. Emerg. Med. Clin. North Am. 33, 133–151. https://doi.org/10.1016/j.emc.2014.09.010

Kirsch, R., Gramzow, L., Theißen, G., Siegfried, B.D., ffrench-Constant, R.H., Heckel, D.G., Pauchet, Y., 2014. Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. Insect Biochem. Mol. Biol. 52, 33–50. https://doi.org/10.1016/j.ibmb.2014.06.008

Kirsch, R., Okamura, Y., Haeger, W., Vogel, H., Kunert, G., Pauchet, Y., 2022. Metabolic novelty originating from horizontal gene transfer is essential for leaf beetle survival. Proc. Natl. Acad. Sci. 119, e2205857119. https://doi.org/10.1073/pnas.2205857119

Koutsovoulos, G.D., Granjeon Noriot, S., Bailly-Bechet, M., Danchin, E.G.J., Rancurel, C., 2022. AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. PLOS Comput. Biol. 18, e1010686. https://doi.org/10.1371/journal.pcbi.1010686

Krenek, P., Samajova, O., Luptovciak, I., Doskocilova, A., Komis, G., Samaj, J., 2015. Transient plant transformation mediated by Agrobacterium tumefaciens: Principles, methods and applications. Biotechnol. Adv. 33, 1024–1042. https://doi.org/10.1016/j.biotechadv.2015.03.012

Kudla, U., Milac, A.-L., Qin, L., Overmars, H., Roze, E., Holterman, M., Petrescu, A.-J., Goverse, A., Bakker, J., Helder, J., Smant, G., 2007. Structural and functional characterization of a novel, host penetration-related pectate lyase from the potato cyst nematode Globodera rostochiensis. Mol. Plant Pathol. 8, 293–305. https://doi.org/10.1111/j.1364-3703.2007.00394.x

Kumar, V., Khan, M.R., Walia, R.K., 2020. Crop Loss Estimations due to Plant-Parasitic Nematodes in Major Crops in India. Natl. Acad. Sci. Lett. 43, 409–412. https://doi.org/10.1007/s40009-020-00895-2

Kundu, P., Mondal, S., Ghosh, A., 2022. Bacterial species metabolic interaction network for deciphering the lignocellulolytic system in fungal cultivating termite gut microbiota. Biosystems 221, 104763. https://doi.org/10.1016/j.biosystems.2022.104763

Lai, C.-K., Lee, Y., Ke, H.-M., Lu, M.R., Liu, W.-A., Lee, H.-H., Liu, Y.-C., Yoshiga, T., Kikuchi, T., Chen, P.J., Tsai, I.J., 2022. The *Aphelenchoides* genomes reveal major events of horizontal gene transfers in clade IV nematodes (preprint). Genomics. https://doi.org/10.1101/2022.09.13.507733

Lake, J.A., Henderson, E., Oakes, M., Clark, M.W., 1984. Eocytes: a new ribosome structure

indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 81, 3786–3790. https://doi.org/10.1073/pnas.81.12.3786

Lambert, K.N., Allen, K.D., Sussex, I.M., 1999. Cloning and Characterization of an Esophageal-Gland-Specific Chorismate Mutase from the Phytoparasitic Nematode *Meloidogyne javanica*. Mol. Plant-Microbe Interactions® 12, 328–336. https://doi.org/10.1094/MPMI.1999.12.4.328

Landrigan, P.J., Stegeman, J.J., Fleming, L.E., Allemand, D., Anderson, D.M., Backer, L.C., Brucker-Davis, F., Chevalier, N., Corra, L., Czerucka, D., Bottein, M.-Y.D., Demeneix, B., Depledge, M., Deheyn, D.D., Dorman, C.J., Fénichel, P., Fisher, S., Gaill, F., Galgani, F., Gaze, W.H., Giuliano, L., Grandjean, P., Hahn, M.E., Hamdoun, A., Hess, P., Judson, B., Laborde, A., McGlade, J., Mu, J., Mustapha, A., Neira, M., Noble, R.T., Pedrotti, M.L., Reddy, C., Rocklöv, J., Scharler, U.M., Shanmugam, H., Taghian, G., Van de Water, J.A.J.M., Vezzulli, L., Weihe, P., Zeka, A., Raps, H., Rampal, P., 2020. Human Health and Ocean Pollution. Ann. Glob. Health 86, 151. https://doi.org/10.5334/aogh.2831

Lanna, A., Vaz, B., D'Ambra, C., Valvo, S., Vuotto, C., Chiurchiù, V., Devine, O., Sanchez, M., Borsellino, G., Akbar, A.N., De Bardi, M., Gilroy, D.W., Dustin, M.L., Blumer, B., Karin, M., 2022. An intercellular transfer of telomeres rescues T cells from senescence and promotes long-term immunological memory. Nat. Cell Biol. https://doi.org/10.1038/s41556-022-00991-z

Lapadula, W.J., Mascotti, M.L., Juri Ayub, M., 2020. Whitefly genomes contain ribotoxin coding genes acquired from plants. Sci. Rep. 10, 15503. https://doi.org/10.1038/s41598-020-72267-1

Lapadula, W.J., Sánchez Puerta, M.V., Juri Ayub, M., 2013. Revising the Taxonomic Distribution, Origin and Evolution of Ribosome Inactivating Protein Genes. PLoS ONE 8, e72825. https://doi.org/10.1371/journal.pone.0072825

Lederberg, J., Tatum, E.L., 1946. Gene Recombination in Escherichia Coli. Nature 158, 558–558. https://doi.org/10.1038/158558a0

Ledger, T.N., Jaubert, S., Bosselut, N., Abad, P., Rosso, M.-N., 2006. Characterization of a new β-1,4-endoglucanase gene from the root-knot nematode Meloidogyne incognita and evolutionary scheme for phytonematode family 5 glycosyl hydrolases. Gene 382, 121–128. https://doi.org/10.1016/j.gene.2006.06.023

Levasseur, A., Orlando, L., Bailly, X., Milinkovitch, M.C., Danchin, E.G.J., Pontarotti, P., 2007. Conceptual bases for quantifying the role of the environment on gene evolution: the participation of positive selection and neutral evolution. Biol. Rev. 82, 551–572. https://doi.org/10.1111/j.1469-185X.2007.00024.x

Li, F.-W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., Burge, D.O., Pokorny, L., Larsson, A., Chen, T., Weststrand, S., Thomas, P., Carpenter, E., Zhang, Y., Tian, Z., Chen, L., Yan, Z., Zhu, Y., Sun, X., Wang, J., Stevenson, D.W., Crandall-Stotler, B.J., Shaw, A.J., Deyholos, M.K., Soltis, D.E., Graham, S.W., Windham, M.D., Langdale, J.A., Wong, G.K.-S., Mathews, S., Pryer, K.M., 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc. Natl. Acad. Sci. 111, 6672–6677. https://doi.org/10.1073/pnas.1319929111

Lilley, C.J., Maqbool, A., Wu, D., Yusup, H.B., Jones, L.M., Birch, P.R.J., Banfield, M.J., Urwin, P.E., Eves-van den Akker, S., 2018. Effector gene birth in plant parasitic nematodes: Neofunctionalization of a housekeeping glutathione synthetase gene. PLOS Genet. 14, e1007310. https://doi.org/10.1371/journal.pgen.1007310

Liu, C., Liu, Z., Fang, Y., Du, Z., Yan, Z., Yuan, X., Dai, L., Yu, T., Xiong, M., Tian, Y., Li, H., Li, F., Zhang, J., Meng, L., Wang, Z., Jiang, H., Zhang, Z., 2022. Exposure to the environmentally toxic pesticide maneb induces Parkinson's disease-like neurotoxicity in mice: A combined proteomic and metabolomic analysis. Chemosphere 308, 136344. https://doi.org/10.1016/j.chemosphere.2022.136344

Liu, J., Berry, R.E., Moldenke, A.F., 1997. Phylogenetic Relationships of Entomopathogenic Nematodes (Heterorhabditidae and Steinernematidae) Inferred from Partial 18S rRNA Gene Sequences. J. Invertebr. Pathol. 69, 246–252. https://doi.org/10.1006/jipa.1997.4657

Loiseau, V., Peccoud, J., Bouzar, C., Guillier, S., Fan, J., Gueli Alletti, G., Meignin, C., Herniou, E.A., Federici, B.A., Wennmann, J.T., Jehle, J.A., Cordaux, R., Gilbert, C., 2021. Monitoring Insect Transposable Elements in Large Double-Stranded DNA Viruses Reveals Host-to-Virus and Virus-to-Virus Transposition. Mol. Biol. Evol. 38, 3512–3530. https://doi.org/10.1093/molbev/msab198

Lucic, B., de Castro, I.J., Lusic, M., 2021. Viruses in the Nucleus. Cold Spring Harb. Perspect. Biol. 13, a039446. https://doi.org/10.1101/cshperspect.a039446

Marchi, N., Winkelbach, L., Schulz, I., Brami, M., Hofmanová, Z., Blöcher, J., Reyna-Blanco, C.S., Diekmann, Y., Thiéry, A., Kapopoulou, A., Link, V., Piuz, V., Kreutzer, S., Figarska, S.M., Ganiatsou, E., Pukaj, A., Struck, T.J., Gutenkunst, R.N., Karul, N., Gerritsen, F., Pechtl, J., Peters, J., Zeeb-Lanz, A., Lenneis, E., Teschler-Nicola, M., Triantaphyllou, S., Stefanović, S., Papageorgopoulou, C., Wegmann, D., Burger, J., Excoffier, L., 2022. The genomic origins of the world's first farmers. Cell 185, 1842-1859.e18. https://doi.org/10.1016/j.cell.2022.04.008

Margulis, L., 1990. Words as Battle Cries: Symbiogenesis and the New Field of Endocytobiology. BioScience 40, 673. https://doi.org/10.2307/1311435

Margulis, L., 1975. Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. Symp. Soc. Exp. Biol. 21–38.

Martin, W.F., 2017. Too Much Eukaryote LGT. BioEssays 39, 1700115.

https://doi.org/10.1002/bies.201700115

Martin, W.F., Garg, S., Zimorski, V., 2015. Endosymbiotic theories for eukaryote origin. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140330. https://doi.org/10.1098/rstb.2014.0330

Matson, P.A., Parton, W.J., Power, A.G., Swift, M.J., 1997. Agricultural Intensification and Ecosystem Properties. Science 277, 504–509. https://doi.org/10.1126/science.277.5325.504

McCarter, J.P., Dautova Mitreva, M., Martin, J., Dante, M., Wylie, T., Rao, U., Pape, D., Bowers, Y., Theising, B., Murphy, C.V., Kloek, A.P., Chiapelli, B.J., Clifton, S.W., Bird, D.M., Waterston, R.H., 2003. Analysis and functional classification of transcripts from the nematode Meloidogyne incognita. Genome Biol. 4, R26. https://doi.org/10.1186/gb-2003-4-4-r26

McNally, A., Thomson, N.R., Reuter, S., Wren, B.W., 2016. "Add, stir and reduce": Yersinia spp. as model bacteria for pathogen evolution. Nat. Rev. Microbiol. 14, 177–190. https://doi.org/10.1038/nrmicro.2015.29

Milot, E., Belmaaza, A., Wallenburg, J.C., Gusew, N., Bradley, W.E., Chartrand, P., 1992. Chromosomal illegitimate recombination in mammalian cells is associated with intrinsically bent DNA elements. EMBO J. 11, 5063–5070. https://doi.org/10.1002/j.1460-2075.1992.tb05613.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. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., 2019. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. gkz1035. https://doi.org/10.1093/nar/gkz1035

Mitchum, M.G., Hussey, R.S., Baum, T.J., Wang, X., Elling, A.A., Wubben, M., Davis, E.L., 2013. Nematode effector proteins: an emerging paradigm of parasitism. New Phytol. 199, 879–894. https://doi.org/10.1111/nph.12323

Mitreva-Dautova, M., Roze, E., Overmars, H., de Graaff, L., Schots, A., Helder, J., Goverse, A., Bakker, J., Smant, G., 2006. A Symbiont-Independent Endo-1,4-β-Xylanase from the Plant-Parasitic Nematode *Meloidogyne incognita*. Mol. Plant-Microbe Interactions® 19, 521–529. https://doi.org/10.1094/MPMI-19-0521

Miyara, S.B., Ionit, I., Buki, P., Kolomiets, M., 2015. The Role of Lipid Signalling in Regulating Plant–Nematode Interactions, in: Advances in Botanical Research. Elsevier, pp. 139–166. https://doi.org/10.1016/bs.abr.2014.12.004

Mohammad Rahimi, H., Yadegar, A., Asadzadeh Aghdaei, H., Mirjalali, H., Zali, M.R., 2022.

Modulation of microRNAs and claudin-7 in Caco-2 cell line treated with Blastocystis sp., subtype 3 soluble total antigen. BMC Microbiol. 22, 111. https://doi.org/10.1186/s12866-022-02528-8

Moran, N.A., Jarvik, T., 2010. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. Science 328, 624–627. https://doi.org/10.1126/science.1187113

Murray, J.A.H., 1987. Micro Review Bending the rules: the 2? plasmid of yeast. Mol. Microbiol. 1, 1–4. https://doi.org/10.1111/j.1365-2958.1987.tb00519.x

Naylor, D., Fansler, S., Brislawn, C., Nelson, W.C., Hofmockel, K.S., Jansson, J.K., McClure, R., 2020. Deconstructing the Soil Microbiome into Reduced-Complexity Functional Modules. mBio 11. https://doi.org/10.1128/mBio.01349-20

Nedorost, Ľ., Pokluda, R., 2012. Effect of arbuscular mycorrhizal fungi on tomato yield and nutrient uptake under different fertilization levels. Acta Univ. Agric. Silvic. Mendel. Brun. 60, 181–186. https://doi.org/10.11118/actaun201260080181

Nicholls, S.M., Quick, J.C., Tang, S., Loman, N.J., 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience 8, giz043. https://doi.org/10.1093/gigascience/giz043

Nicol, J.M., Turner, S.J., Coyne, D.L., Nijs, L. den, Hockland, S., Maafi, Z.T., 2011. Current Nematode Threats to World Agriculture, in: Jones, J., Gheysen, G., Fenoll, C. (Eds.), Genomics and Molecular Genetics of Plant-Nematode Interactions. Springer Netherlands, pp. 21–43. https://doi.org/10.1007/978-94-007-0434-3_2

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.-L., Wincker, P., Casaregola, S., Dequin, S., 2009. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proc. Natl. Acad. Sci. 106, 16333–16338. https://doi.org/10.1073/pnas.0904673106

Ochman, H., Davalos, L.M., 2006. The Nature and Dynamics of Bacterial Genomes. Science 311, 1730–1733. https://doi.org/10.1126/science.1119966

Ohno, S., 2013. Evolution by Gene Duplication. Springer Science & Business Media.

Okulewicz, A., 2017. The impact of global climate change on the spread of parasitic nematodes. Ann. Parasitol. 15–20. https://doi.org/10.17420/ap6301.79

Onozawa, M., Zhang, Z., Kim, Y.J., Goldberg, L., Varga, T., Bergsagel, P.L., Kuehl, W.M., Aplan, P.D., 2014. Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. Proc. Natl. Acad. Sci. 111, 7729–7734. https://doi.org/10.1073/pnas.1321889111

Opik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J.M., Reier, U., Zobel, M., 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytol. 188, 223–241. https://doi.org/10.1111/j.1469-8137.2010.03334.x

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P., Windham, E., 2008. Sequence and genetic map of *Meloidogyne hapla* : A compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. 105, 14802–14807. https://doi.org/10.1073/pnas.0805946105

Osborn, H.F., 1902. The Fossil Tree Bridge in the Arizona Petrified Forest. Science 16, 991–991. https://doi.org/10.1126/science.16.416.991-b

Paez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., Huntemann, M., K.?Reddy, T.B., Pavlopoulos, G.A., Sullivan, M.B., Campbell, B.J., Chen, F., McMahon, K., Hallam, S.J., Denef, V., Cavicchioli, R., Caffrey, S.M., Streit, W.R., Webster, J., Handley, K.M., Salekdeh, G.H., Tsesmetzis, N., Setubal, J.C., Pope, P.B., Liu, W.-T., Rivers, A.R., Ivanova, N.N., Kyrpides, N.C., 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 45, D457–D465. https://doi.org/10.1093/nar/gkw1030

Paganini, J., Campan-Fournier, A., Da Rocha, M., Gouret, P., Pontarotti, P., Wajnberg, E., Abad, P., Danchin, E.G.J., 2012. Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. PLoS ONE 7, e50875. https://doi.org/10.1371/journal.pone.0050875

Palazzo, A.F., Lee, E.S., 2015. Non-coding RNA: what is functional and what is junk? Front. Genet. 6. https://doi.org/10.3389/fgene.2015.00002

Palomares-Rius, J.E., Hirooka, Y., Tsai, I.J., Masuya, H., Hino, A., Kanzaki, N., Jones, J.T., Kikuchi, T., 2014. Distribution and evolution of glycoside hydrolase family 45 cellulases in nematodes and fungi. BMC Evol. Biol. 14, 69. https://doi.org/10.1186/1471-2148-14-69

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50, D785–D794. https://doi.org/10.1093/nar/gkab776

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pasteur, M., 1884. Infectious Diseases and Vaccination for Rabies. Atlanta Med. Surg. J. 1884 1, 437–448.

Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A.V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., Briggs, J.A.G., Feschotte, C., Shepherd, J.D., 2018. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell 172, 275-288.e18. https://doi.org/10.1016/j.cell.2017.12.024

Pauchet, Y., Wilkinson, P., Chauhan, R., ffrench-Constant, R.H., 2010. Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. PLoS ONE 5, e15635. https://doi.org/10.1371/journal.pone.0015635

Perry, R.N., Moens, M., Starr, J.L. (Eds.), 2009. Root-knot nematodes. CABI North American Office, Cambridge, MA.

Pienaar, R.D., Gilbert, C., Belliardo, C., Herrero, S., Herniou, E.A., 2022. First Evidence of Past and Present Interactions between Viruses and the Black Soldier Fly, Hermetia illucens. Viruses 14, 1274. https://doi.org/10.3390/v14061274

Pierce, null, Biron, null, Rumpho, null, 1996. Endosymbiotic chloroplasts in molluscan cells contain proteins synthesized after plastid capture. J. Exp. Biol. 199, 2323–2330. https://doi.org/10.1242/jeb.199.10.2323

Pierce, S.K., Fang, X., Schwartz, J.A., Jiang, X., Zhao, W., Curtis, N.E., Kocot, K.M., Yang, B., Wang, J., 2012. Transcriptomic evidence for the expression of horizontally transferred algal nuclear genes in the photosynthetic sea slug, Elysia chlorotica. Mol. Biol. Evol. 29, 1545–1556. https://doi.org/10.1093/molbev/msr316

Ponce-Toledo, R.I., López-García, P., Moreira, D., 2019. Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol. 224, 618–624. https://doi.org/10.1111/nph.15965

Popeijus, H., Overmars, H., Jones, J., Blok, V., Goverse, A., Helder, J., Schots, A., Bakker, J., Smant, G., 2000. Degradation of plant cell walls by a nematode. Nature 406, 36–37. https://doi.org/10.1038/35017641

Qin, L., Kudla, U., Roze, E.H.A., Goverse, A., Popeijus, H., Nieuwland, J., Overmars, H., Jones, J.T., Schots, A., Smant, G., Bakker, J., Helder, J., 2004. A nematode expansin acting on plants. Nature 427, 30–30. https://doi.org/10.1038/427030a

Quist, C.W., Smant, G., Helder, J., 2015. Evolution of Plant Parasitism in the Phylum Nematoda. Annu. Rev. Phytopathol. 53, 289–310. https://doi.org/10.1146/annurev-phyto-080614-120057

Rahman, K., Zhang, D., 2018. Effects of Fertilizer Broadcasting on the Excessive Use of Inorganic Fertilizers and Environmental Sustainability. Sustainability 10, 759. https://doi.org/10.3390/su10030759

Rancurel, C., Legrand, L., Danchin, E., 2017. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. Genes 8, 248. https://doi.org/10.3390/genes8100248

Remaut, K., Symens, N., Lucas, B., Demeester, J., De Smedt, S.C., 2014. Cell division responsive peptides for optimized plasmid DNA delivery: The mitotic window of opportunity? J. Controlled Release 179, 1–9. https://doi.org/10.1016/j.jconrel.2014.01.013

Rivera, M.C., Lake, J.A., 1992. Evidence that eukaryotes and eocyte prokaryotes are

immediate relatives. Science 257, 74–76. https://doi.org/10.1126/science.1621096

Roberts, D.P., Denny, T.P., Schell, M.A., 1988. Cloning of the egl gene of Pseudomonas solanacearum and analysis of its role in phytopathogenicity. J. Bacteriol. 170, 1445–1451. https://doi.org/10.1128/jb.170.4.1445-1451.1988

Rosenheim, J.A., Cluff, E., Lippey, M.K., Cass, B.N., Paredes, D., Parsa, S., Karp, D.S., Chaplin-Kramer, R., 2022. Increasing crop field size does not consistently exacerbate insect pest problems. Proc. Natl. Acad. Sci. 119, e2208813119. https://doi.org/10.1073/pnas.2208813119

Ross, P., Mayer, R., Benziman, M., 1991. Cellulose biosynthesis and function in bacteria. Microbiol. Rev. 55, 35–58. https://doi.org/10.1128/mr.55.1.35-58.1991

Rosso, M.-N., Favery, B., Piotte, C., Arthaud, L., De Boer, J.M., Hussey, R.S., Bakker, J., Baum, T.J., Abad, P., 1999a. Isolation of a cDNA Encoding a β-1,4-endoglucanase in the Root-Knot Nematode *Meloidogyne incognita* and Expression Analysis During Plant Parasitism. Mol. Plant-Microbe Interactions® 12, 585–591. https://doi.org/10.1094/MPMI.1999.12.7.585

Rosso, M.-N., Favery, B., Piotte, C., Arthaud, L., De Boer, J.M., Hussey, R.S., Bakker, J., Baum, T.J., Abad, P., 1999b. Isolation of a cDNA Encoding a β-1,4-endoglucanase in the Root-Knot Nematode *Meloidogyne incognita* and Expression Analysis During Plant Parasitism. Mol. Plant-Microbe Interactions® 12, 585–591. https://doi.org/10.1094/MPMI.1999.12.7.585

Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A., Manhart, J.R., 2008. Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug Elysia chlorotica. Proc. Natl. Acad. Sci. U. S. A. 105, 17867–17871. https://doi.org/10.1073/pnas.0804968105

Rush, M.G., Misra, R., 1985. Extrachromosomal DNA in eucaryotes. Plasmid 14, 177–191. https://doi.org/10.1016/0147-619X(85)90001-0

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nat. Ecol. Evol. 3, 430–439. https://doi.org/10.1038/s41559-018-0793-y

Scheu, S., Ruess, L., Bonkowski, M., 2005. Interactions Between Microorganisms and Soil Micro- and Mesofauna, in: Varma, A., Buscot, F. (Eds.), Microorganisms in Soils: Roles in Genesis and Functions, Soil Biology. Springer-Verlag, Berlin/Heidelberg, pp. 253–275. https://doi.org/10.1007/3-540-26609-7_12

Scholl, E.H., Thorne, J.L., McCarter, J.P., Bird, D.M., 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. 4, R39. https://doi.org/10.1186/gb-2003-4-6-r39

Schouteden, N., De Waele, D., Panis, B., Vos, C.M., 2015. Arbuscular Mycorrhizal Fungi for the Biocontrol of Plant-Parasitic Nematodes: A Review of the Mechanisms Involved.

Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.01280

Seenivasagan, R., Babalola, O.O., 2021. Utilization of Microbial Consortia as Biofertilizers and Biopesticides for the Production of Feasible Agricultural Product. Biology 10, 1111. https://doi.org/10.3390/biology10111111

Sellitto, V.M., Golubkina, N.A., Pietrantonio, L., Cozzolino, E., Cuciniello, A., Cenvinzo, V., Florin, I., Caruso, G., 2019. Tomato Yield, Quality, Mineral Composition and Antioxidants as Affected by Beneficial Microorganisms Under Soil Salinity Induced by Balanced Nutrient Solutions. Agriculture 9, 110. https://doi.org/10.3390/agriculture9050110

Sermonti, G., Spada-Sermonti, I., 1955. Genetic Recombination in Streptomyces. Nature 176, 121–121. https://doi.org/10.1038/176121a0

Shin, N.R., Doucet, D., Pauchet, Y., 2022. Duplication of Horizontally Acquired GH5_2 Enzymes Played a Central Role in the Evolution of Longhorned Beetles. Mol. Biol. Evol. 39, msac128. https://doi.org/10.1093/molbev/msac128

Sibbald, S.J., Eme, L., Archibald, J.M., Roger, A.J., 2020. Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. Trends Parasitol. 36, 927–941. https://doi.org/10.1016/j.pt.2020.07.014

Singh, S., Singh, B., Singh, A.P., 2015. Nematodes: A Threat to Sustainability of Agriculture. Procedia Environ. Sci. 29, 215–216. https://doi.org/10.1016/j.proenv.2015.07.270

Smant, G., Stokkermans, J.P.W.G., Yan, Y., de Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L., Helder, J., Schots, A., Bakker, J., 1998. Endogenous cellulases in animals: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc. Natl. Acad. Sci. 95, 4906–4911. https://doi.org/10.1073/pnas.95.9.4906

Soler, N., Forterre, P., 2020. Vesiduction: the fourth way of HGT. Environ. Microbiol. 22, 2457–2460. https://doi.org/10.1111/1462-2920.15056

Soni, V.K., Krishnapriya, R., Sharma, R.K., 2021. Algae: Biomass to Biofuel, in: Basu, C. (Ed.), Biofuels and Biodiesel, Methods in Molecular Biology. Springer US, New York, NY, pp. 31–51. https://doi.org/10.1007/978-1-0716-1323-8_3

Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L., Ettema, T.J.G., 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521, 173–179. https://doi.org/10.1038/nature14447

Sponsler, O.L., 1923. STRUCTURAL UNITS OF STARCH DETERMINED BY X-RAY CRYSTAL STRUCTURE METHOD. J. Gen. Physiol. 5, 757–776. https://doi.org/10.1085/jgp.5.6.757

Stecher, B., Maier, L., Hardt, W.-D., 2013. "Blooming" in the gut: how dysbiosis might contribute to pathogen evolution. Nat. Rev. Microbiol. 11, 277–284.

https://doi.org/10.1038/nrmicro2989

Subbotin, S.A., Palomares Rius, J.E., Castillo, P., 2021. Systematics of Root-knot Nematodes (Nematoda: Meloidogynidae). BRILL. https://doi.org/10.1163/9789004387584

Syvanen, M., 1987. Molecular clocks and evolutionary relationships: Possible distortions due to horizontal gene flow. J. Mol. Evol. 26, 16–23. https://doi.org/10.1007/BF02111278

Syvanen, M., 1985. Cross-species gene transfer; implications for a new theory of evolution. J. Theor. Biol. 112, 333–343. https://doi.org/10.1016/S0022-5193(85)80291-5

Szakasits, D., Heinen, P., Wieczorek, K., Hofmann, J., Wagner, F., Kreil, D.P., Sykacek, P., Grundler, F.M.W., Bohlmann, H., 2009. The transcriptome of syncytia induced by the cyst nematode *Heterodera schachtii* in Arabidopsis roots. Plant J. 57, 771–784. https://doi.org/10.1111/j.1365-313X.2008.03727.x

Taoheed, A.M., Ateka, E.M., Losenge, T., 2018. ARBUSCULAR MYCORRHIZA FUNGI PROMOTES GROWTH OF TOMATO SEEDLINGS IN THE ABSENCE OF PHOSPHATE IN NUTRIENT SOLUTION 7, 9.

Taylor, C.E., Robertson, W.M., 1970. Sites of virus retention in the alimentary tract of the nematode vectors, Xiphinema diversicaudatum (Micol.) and X. index (Thorne and Allen). Ann. Appl. Biol. 66, 375–380. https://doi.org/10.1111/j.1744-7348.1970.tb04616.x

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., Smith, M.E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Põldmaa, K., Piepenbring, M., Phosri, C., Peterson, M., Parts, K., Pärtel, K., Otsing, E., Nouhra, E., Njouonkou, A.L., Nilsson, R.H., Morgado, L.N., Mayor, J., May, T.W., Majuakim, L., Lodge, D.J., Lee, S.S., Larsson, K.-H., Kohout, P., Hosaka, K., Hiiesalu, I., Henkel, T.W., Harend, H., Guo, L., Greslebin, A., Grelet, G., Geml, J., Gates, G., Dunstan, W., Dunk, C., Drenkhan, R., Dearnaley, J., De Kesel, A., Dang, T., Chen, X., Buegger, F., Brearley, F.Q., Bonito, G., Anslan, S., Abell, S., Abarenkov, K., 2014. Global diversity and geography of soil fungi. Science 346, 1256688. https://doi.org/10.1126/science.1256688

Thomas, C.M., Nielsen, K.M., 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat. Rev. Microbiol. 3, 711–721. https://doi.org/10.1038/nrmicro1234

Toju, H., Sato, H., Yamamoto, S., Tanabe, A.S., 2018. Structural diversity across arbuscular mycorrhizal, ectomycorrhizal, and endophytic plant–fungus networks. BMC Plant Biol. 18, 292. https://doi.org/10.1186/s12870-018-1500-5

Tomimatsu, Y., Scherer, J.R., Yeh, Y., Feeney, R.E., 1976. Raman spectra of a solid antifreeze glycoprotein and its liquid and frozen aqueous solutions. J. Biol. Chem. 251, 2290–2298. https://doi.org/10.1016/S0021-9258(17)33585-8

Topalović, O., Santos, S.S., Heuer, H., Nesme, J., Kanfra, X., Hallmann, J., Sørensen, S.J.,

Vestergård, M., 2022. Deciphering bacteria associated with a pre-parasitic stage of the root-knot nematode Meloidogyne hapla in nemato-suppressive and nemato-conducive soils. Appl. Soil Ecol. 172, 104344. https://doi.org/10.1016/j.apsoil.2021.104344

Torsvik, V., Øvreås, L., 2002. Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. 5, 240–245. https://doi.org/10.1016/S1369-5274(02)00324-7

Trench, R.K., 1975. Of "leaves that crawl": functional chloroplasts in animal cells. Symp. Soc. Exp. Biol. 229–265.

True, J.R., Carroll, S.B., 2002. Gene Co-Option in Physiological and Morphological Evolution. Annu. Rev. Cell Dev. Biol. 18, 53–80. https://doi.org/10.1146/annurev.cellbio.18.020402.140619

Tscharntke, T., Klein, A.M., Kruess, A., Steffan-Dewenter, I., Thies, C., 2005. Landscape perspectives on agricultural intensification and biodiversity – ecosystem service management. Ecol. Lett. 8, 857–874. https://doi.org/10.1111/j.1461-0248.2005.00782.x

Van Den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D.A., de Goede, R.G.M., Adams, B.J., Ahmad, W., Andriuzzi, W.S., Bardgett, R.D., Bonkowski, M., Campos-Herrera, R., Cares, J.E., Caruso, T., de Brito Caixeta, L., Chen, X., Costa, S.R., Creamer, R., Mauro da Cunha Castro, J., Dam, M., Djigal, D., Escuer, M., Griffiths, B.S., Gutiérrez, C., Hohberg, K., Kalinkina, D., Kardol, P., Kergunteuil, A., Korthals, G., Krashevska, V., Kudrin, A.A., Li, Q., Liang, W., Magilton, M., Marais, M., Martín, J.A.R., Matveeva, E., Mayad, E.H., Mulder, C., Mullin, P., Neilson, R., Nguyen, T.A.D., Nielsen, U.N., Okada, H., Rius, J.E.P., Pan, K., Peneva, V., Pellissier, L., Carlos Pereira da Silva, J., Pitteloud, C., Powers, T.O., Powers, K., Quist, C.W., Rasmann, S., Moreno, S.S., Scheu, S., Setälä, H., Sushchuk, A., Tiunov, A.V., Trap, J., van der Putten, W., Vestergård, M., Villenave, C., Waeyenberge, L., Wall, D.H., Wilschut, R., Wright, D.G., Yang, J., Crowther, T.W., 2019. Soil nematode abundance and functional group composition at a global scale. Nature 572, 194–198. https://doi.org/10.1038/s41586-019-1418-6

van Megen, H., van den Elsen, S., Holterman, M., Karssen, G., Mooyman, P., Bongers, T., Holovachov, O., Bakker, J., Helder, J., 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology 11, 927–950. https://doi.org/10.1163/156854109X456862

Van Oss, S.B., Carvunis, A.-R., 2019. De novo gene birth. PLOS Genet. 15, e1008160. https://doi.org/10.1371/journal.pgen.1008160

Vanholme, B., Kast, P., Haegeman, A., Jacob, J., Grunewald, W., Gheysen, G., 2009. Structural and functional investigation of a secreted chorismate mutase from the plant-parasitic nematode *Heterodera schachtii* in the context of related enzymes from diverse origins. Mol. Plant Pathol. 10, 189–200. https://doi.org/10.1111/j.1364-3703.2008.00521.x

Veronico, P., Jones, J., Di Vito, M., De Giorgi, C., 2001. Horizontal transfer of a bacterial gene

involved in polyglutamate biosynthesis to the plant-parasitic nematode *Meloidogyne artiellia*. FEBS Lett. 508, 470–474. https://doi.org/10.1016/S0014-5793(01)03132-5

Vieira, P., Shao, J., Vijayapalani, P., Maier, T.R., Pellegrin, C., Eves-van den Akker, S., Baum, T.J., Nemchinov, L.G., 2020. A new esophageal gland transcriptome reveals signatures of large scale de novo effector birth in the root lesion nematode Pratylenchus penetrans. BMC Genomics 21, 738. https://doi.org/10.1186/s12864-020-07146-0

Vollset, S.E., Goren, E., Yuan, C.-W., Cao, J., Smith, A.E., Hsiao, T., Bisignano, C., Azhar, G.S., Castro, E., Chalek, J., Dolgert, A.J., Frank, T., Fukutaki, K., Hay, S.I., Lozano, R., Mokdad, A.H., Nandakumar, V., Pierce, M., Pletcher, M., Robalik, T., Steuben, K.M., Wunrow, H.Y., Zlavog, B.S., Murray, C.J.L., 2020. Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. The Lancet 396, 1285–1306. https://doi.org/10.1016/S0140-6736(20)30677-2

Wagner, A., Whitaker, R.J., Krause, D.J., Heilers, J.-H., van Wolferen, M., van der Does, C., Albers, S.-V., 2017. Mechanisms of gene flow in archaea. Nat. Rev. Microbiol. 15, 492–501. https://doi.org/10.1038/nrmicro.2017.41

Wang, L., Deng, D., Feng, Q., Xu, Z., Pan, H., Li, H., 2022. Changes in litter input exert divergent effects on the soil microbial community and function in stands of different densities. Sci. Total Environ. 845, 157297. https://doi.org/10.1016/j.scitotenv.2022.157297

Wasmuth, J., Schmid, R., Hedley, A., Blaxter, M., 2008. On the Extent and Origins of Genic Novelty in the Phylum Nematoda. PLoS Negl. Trop. Dis. 2, e258. https://doi.org/10.1371/journal.pntd.0000258

Watanabe, T., 1963. INFECTIVE HEREDITY OF MULTIPLE DRUG RESISTANCE IN BACTERIA. Bacteriol. Rev. 27, 87–115. https://doi.org/10.1128/br.27.1.87-115.1963

Wei, W., Schon, K.R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., Tischkowitz, M., Caulfield, M.J., Chinnery, P.F., 2022. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature. https://doi.org/10.1038/s41586-022-05288-7

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Weyandt, N., Aghdam, S.A., Brown, A.M.V., 2022. Discovery of Early-Branching Wolbachia Reveals Functional Enrichment on Horizontally Transferred Genes. Front. Microbiol. 13, 867392. https://doi.org/10.3389/fmicb.2022.867392

Wheeler, T.J., Eddy, S.R., 2013. nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489. https://doi.org/10.1093/bioinformatics/btt403

Whitcomb, J.M., 1992. RETROVIRAL REVERSE TRANSCRIPTION AND INTEGRATION: Progress and Problems. Annu. Rev. Celt Bioi. 8, 275–306.

Willemsen, A., Félez-Sánchez, M., Bravo, I.G., 2019. Genome Plasticity in Papillomaviruses and De Novo Emergence of E5 Oncogenes. Genome Biol. Evol. 11, 1602–1617. https://doi.org/10.1093/gbe/evz095

Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J., Embley, T.M., 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat. Ecol. Evol. 4, 138–147. https://doi.org/10.1038/s41559-019-1040-x

Woehle, C., Roy, A.-S., Glock, N., Wein, T., Weissenbach, J., Rosenstiel, P., Hiebenthal, C., Michels, J., Schönfeld, J., Dagan, T., 2018. A Novel Eukaryotic Denitrification Pathway in Foraminifera. Curr. Biol. CB 28, 2536-2543.e5. https://doi.org/10.1016/j.cub.2018.06.027

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. 87, 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wu, C., Zhao, J., Li, Z., Liu, W., Mei, X., Ning, J., She, D., 2019. Modeling of the Phytophthora capsici cellulose synthase 3 and its inhibitors activity assay. Pest Manag. Sci. 75, 3024–3030. https://doi.org/10.1002/ps.5417

Wybouw, N., Balabanidou, V., Ballhorn, D.J., Dermauw, W., Grbić, M., Vontas, J., Van Leeuwen, T., 2012. A horizontally transferred cyanase gene in the spider mite Tetranychus urticae is involved in cyanate metabolism and is differentially expressed upon host plant change. Insect Biochem. Mol. Biol. 42, 881–889. https://doi.org/10.1016/j.ibmb.2012.08.002

Wybouw, N., Dermauw, W., Tirry, L., Stevens, C., Grbić, M., Feyereisen, R., Van Leeuwen, T., 2014. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. eLife 3, e02365. https://doi.org/10.7554/eLife.02365

Yamasaki, H., Fujimoto, S. & Miyazaki, K. Phylogenetic position of Loricifera inferred from nearly complete 18S and 28S rRNA gene sequences. Zoological Lett 1, 18 (2015). https://doi.org/10.1186/s40851-015-0017-0

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., Stott, M.B., Nunoura, T., Banfield, J.F., Schramm, A., Baker, B.J., Spang, A., Ettema, T.J.G., 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353–358. https://doi.org/10.1038/nature21031

Zarlenga, D., Thompson, P., Mitreva, M., Rosa, B.A., Hoberg, E., 2022. Horizontal gene transfer provides insights into the deep evolutionary history and biology of Trichinella. Food Waterborne Parasitol. 27, e00155. https://doi.org/10.1016/j.fawpar.2022.e00155

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., Gilbert, C., 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat. Commun. 11, 1362. https://doi.org/10.1038/s41467-020-15149-4

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y., Wing, R.A., Liu, S., Long, M., 2019. Rapid evolution of protein diversity by de novo origination in Oryza. Nat. Ecol. Evol. 3, 679–690. https://doi.org/10.1038/s41559-019-0822-5

Zillig, W., Prangishvili, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D., Götz, D., 1996. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic *Archaea*. FEMS Microbiol. Rev. 18, 225–236. https://doi.org/10.1111/j.1574-6976.1996.tb00239.x

Zinder, N.D., Lederberg, J., 1952. GENETIC EXCHANGE IN SALMONELLA. J. Bacteriol. 64, 679–699. https://doi.org/10.1128/jb.64.5.679-699.1952