Rapport de stage

Du 08 janvier 2024 au 28 juin 2024

Évaluation de la qualité de prédiction des annotations de génomes fongiques par approche sur références contenues dans MycoCosm

Mehdi Bourema

Master 2 en biologie structurale et bio-informatique

Faculté des sciences de la vie – Université de Strasbourg

Maître de stage :

Dr. Auer Lucas

Organisme :

Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)

Centre Grand Est - Nancy
Rue d'Amance, 54280 Champenoux

Département Ecologie et biodiversité

UMR Interactions Arbres-Microorganisme

Résumé :

Les champignons sont des organismes eucaryotes qui jouent un rôle majeur dans de nombreux écosystèmes, en particulier dans les cycles biogéochimiques (carbone, nitrogène et éléments minéraux). Ils peuvent être classifiés en guildes écologiques selon les fonctions qu'ils expriment, comme décomposeurs de matière organique, parasites et pathogènes, symbiotiques des plantes. La compréhension de mécanismes résultants de l'interactions des différentes communautés fongiques est cruciale en biologie du sol. Elle peut par exemple aider à identifier des leviers pour améliorer le stockage de carbone dans les sols, ce qui peut être un outil efficace pour atténuer le changement climatique. Cette compréhension implique l'usage d'approches fonctionnelles, notamment au travers d'approches de métatranscriptomiques pour lesquelles identifier les fonctions exprimées par chaque guilde écologique est critique.

La métatranscriptomique des champignons du sol fait face à la très grande biodiversité d'espèces, amenant souvent à une reconstruction incomplète des transcrits, rendant les efforts d'annotation difficiles. L'approche la plus prometteuse pour répondre à cette problématique est l'utilisation d'annotations comparatives basées sur des références contenues dans des bases de données, en particulier MycoCosm qui présente plus de 2500 génomes fongiques en 2024. Cependant, tandis que ces approches ont fourni des résultats solides et prometteurs, aucune évaluation méthodologique de la qualité des prédictions d'annotations n'a encore été menée.

Ce projet vise à remédier à ce manque en évaluant la précision et la fiabilité des résultats d'annotations obtenus par les approches sur références. Cette évaluation requiert une comparaison des annotations attendues par rapport aux annotations prédites. Pour cela une approche basée sur l'exploitation de 2519 génomes annotés dans MycoCosm et consistant à retirer chaque génome de la base de données et à l'interroger contre les génomes restants de manière systématique sur l'ensemble de la phylogénie du règne des Fungi a été adoptée. A partir de l'ensemble des données résultantes, différentes combinaisons de taxonomie, guildes et fonctions ont été étudiés.

Table des matières

Abréviations

OTU : Unité taxonomique opérationnelle (Operational Taxonomic Unit)

KOG : Classification et regroupement des protéines orthologues d'organismes eucaryotes en fonction de leur similarité de séquence et de fonction (EuKaryotic Orthologous Group)

SigP : Courte séquence d'acides aminés présente à l'extrémité N-terminale des protéines destinées à être sécrétées (Signal Peptide)

Closest_ref : Plus proche référence taxonomique (genre, famille ... jusqu'au phylum) des génomes de la base de données (Closest taxonomic reference)

CDS : Séquence codante (Coding Cequence)

AA : Séquence d'acides aminés (Amino Acid)

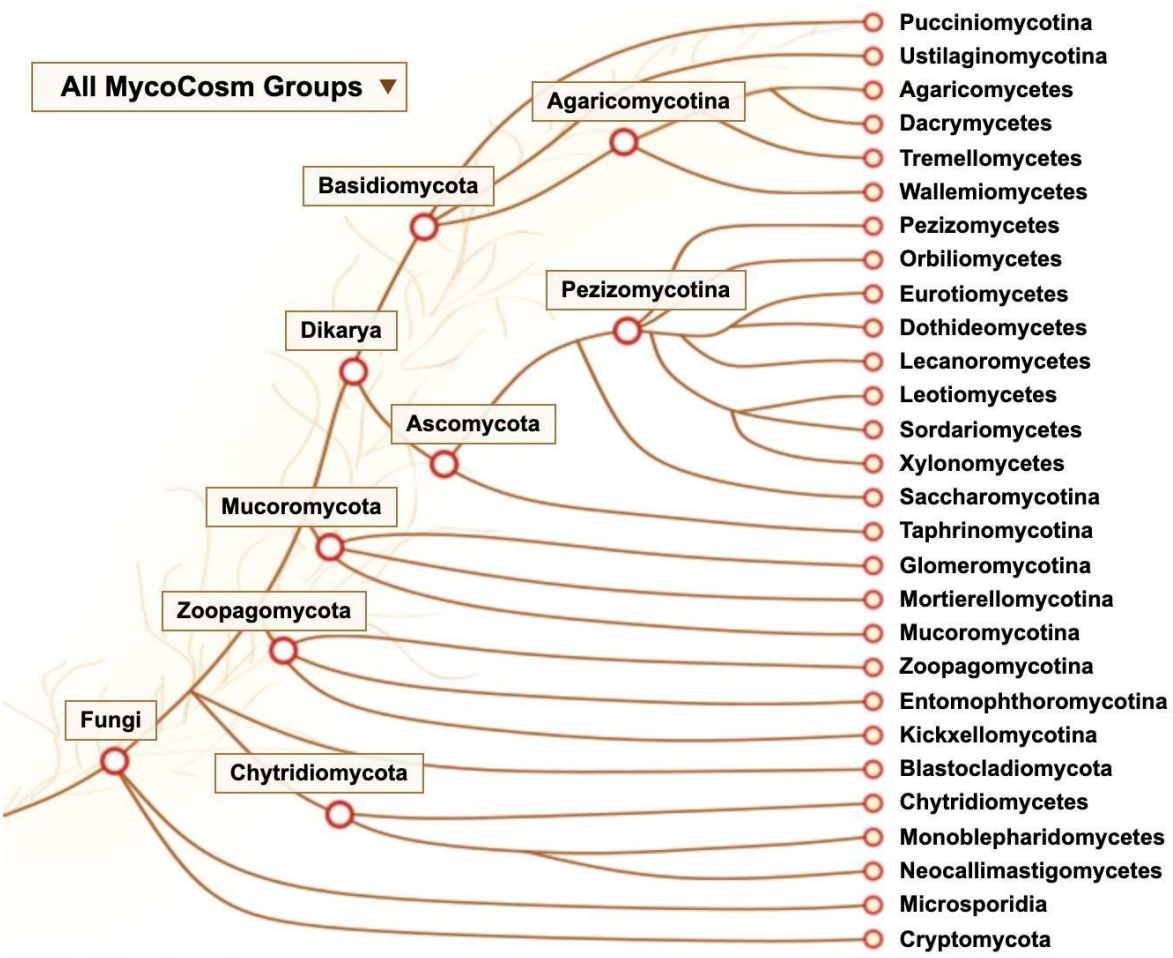

Figure 1 : Arbre phylogénétique du règne Fungi d'après MycoCosm.

Remerciements

Je tiens à remercier dans un premier temps le centre INRAE de Champenoux de m'avoir offert l'opportunité de réaliser ce travail.

Je souhaite aussi exprimer ma profonde gratitude à mon tuteur, M. Lucas Auer, pour le partage de son expertise, sa patience, ses remarques et corrections détaillées tout au long de mon stage.

Enfin, je remercie les personnes qui ont, de près ou de loin, contribués à la réalisation de ce travail.

1. Introduction :

Les champignons sont des organismes eucaryotes répartis dans le monde entier. Ils ont colonisé l'ensemble des niches écologiques disponibles, couvrant un spectre maximal de ressources et d'habitats (Skovgaard, 2002). Les différentes espèces ont des répartitions géographiques distinctes en fonction de leurs hôtes et du climat (Boddy, 2016). On peut ainsi les retrouver dans le sol, l'eau et bien d'autres surfaces, en association avec les animaux, les plantes et d'autres organismes. Ils peuvent aussi bien être des micro-organismes unicellulaires comme les levures que des complexes multicellulaires visibles à l'œil nu comme les champignons macroscopiques. Ils jouent un rôle majeur dans la bioconversion en participant à la continuation des cycles biogéochimiques d'énergie et de nutriments qui sont essentiels à la vie. Leurs applications industrielles ou médicales sont nombreuses et très anciennes, des fermentations alimentaires à la pénicilline. Leur métabolisme est communément exploité dans les industries biotechnologiques et pharmaceutiques pour obtenir de nombreux antibiotiques, polysaccharides, vitamines, lipides, enzymes et d'autres produits inaccessibles à partir d'autres organismes (van der Linde et al., 2018, p. Ectomy). Les champignons peuvent également avoir des effets négatifs : certains d'entre eux représentent une menace infectieuse pour l'Homme et les animaux (Hernandez & Martinez, 2018), mais ils ont surtout un potentiel de dommage très élevé pour l'agriculture, l'alimentation humaine et les végétaux en général (Case et al., 2022). Les pertes de récoltes dues aux pathogènes végétaux, en particulier les champignons, sont estimées à 21 milliards par an aux Etats-Unis, soit plus que les pertes causées par les insectes non indigènes (Rossman, 2009). Les plantes présentent plusieurs niveaux de résistance aux infections fongiques, cependant certains facteurs tels que le stress thermique, les dégâts causés par les insectes et la sécheresse, affaiblissent la résistance des cultures à ces champignons et jouent un rôle important dans l'invasion fongique de ces cultures et dans la production de toxines subséquente. A contrario, certains champignons endophytes peuvent jouer un rôle important dans la croissance et la résistance à différentes maladies, stress biotiques et abiotiques des plantes (Fatma A. Abo Nouh, 2019). Enfin, certaines espèces peuvent causer des infections humaines chez des individus sains et sont de plus en plus les agents responsables d'infection opportunistes mortelles chez les patients immunodéprimés (AJ De Lucca, 2007).

Les Fungi constituent un des clades les plus diversifiés d'eucaryotes dans les écosystèmes terrestres. Dans la classification actuelle ils sont divisés en 8 phylas (figure 1). Plusieurs groupes sont actuellement assez peu diversifiés. Les groupes les plus important et diversifiés sont les *Ascomycota* et les *Basidiomycota*, ils forment le groupe des *Dikaria* (Spatafora et al., 2017).

Ces deux derniers phylas comprennent la majorité des espèces de champignons décrites, et contiennent la plupart des groupes à importance économique : levures, moisissures, pathogènes animaux ou des cultures, symbiotiques.

Ces dernières années, les progrès technologiques ont permis aux chercheurs de séquencer rapidement et à des coûts de plus en plus faibles les micro-organismes présents dans un échantillon environnemental. Cela a conduit à une appréciation croissante de la nature très diversifiée des communautés microbiennes, à la fois à l'échelle locale et mondiale (Tedersoo et al., 2014). En règle générale, les ensembles de données basés sur le séquençage à haut débit contiennent des millions de séquences et des milliers d'unités taxonomiques opérationnelles (OTU). Ce concept qui construit des taxons définis “mathématiquement” est largement accepté et appliqué pour décrire les communautés bactériennes à l’aide du séquençage par amplicon du gène de l’ARNr 16S (Lladó Fernández et al., 2019), mais aussi des champignons en s’appuyant sur le séquençage de l’ADNr 18S ou la région ITS. Les OTU peuvent être utilisées pour déduire des traits fonctionnels quand ceux-ci sont documentées comme étant caractéristiques d’une espèce. Cependant ce lien entre taxons et fonctions a ses limites : les souches d’une même espèce (appartenant idéalement à la même OTU) peuvent ne partager que certains gènes (on parle de génome central), tandis que d’autres gènes sont uniques et spécifiques à la souche. Il est donc difficile de savoir dans quelle mesure les traits fonctionnels importants sont homogènes au sein d’une OTU et dans quelle mesure ils peuvent être déduits pour les membres individuels d’une OTU. Des progrès ont été réalisés pour les procaryotes en reliant des familles de gènes fonctionnels à des groupes phylogénétiques (Langille et al., 2013), mais cette question reste importante pour les organismes eucaryotes tels que les champignons, qui, par rapport aux procaryotes, ont de grands génomes et relativement peu d'espèces entièrement séquencées (Peay, 2014).

Le concept de guilde (également appelé « groupe fonctionnel »), a été créé très tôt en écologie (Schimper et Fisher, 1902) et se réfère globalement à un groupe d'espèces, apparentées ou non, qui exploitent la même catégorie de ressources environnementales de manière similaire (Root, 1967). Les guildes permettent de répartir des communautés taxonomiquement complexes en unités écologiques plus faciles à gérer. Elles offrent également une perspective différente sur la composition des communautés que les mesures basées sur la richesse en espèces ou l'identité taxonomique, en raison de l'accent mis sur les stratégies trophiques. Aussi, l'utilisation de guildes permet de réaliser des études comparatives entre différentes communautés, même lorsqu'il n'y a pas de chevauchement direct dans la composition des espèces (Hawkins et MacMahon, 1989).

Les champignons peuvent être classés en guildes trophiques, les plus importantes dans les écosystèmes forestiers étant les mycorhiziens, les pathogènes et les décomposeurs/saprotrophes.

Les champignons mycorhiziens sont les champignons qui établissent une association symbiotique avec les racines des plantes, en formant un organe mixte plante/champignons appelé mycorhize (Shi et al., 2023). Cette symbiose améliore l'accès à l'eau et aux nutriments minéraux (notamment phosphore et azote) pour la plante qui en échange fournit du carbone photosynthétique (sucres et acides gras) au champignon symbiotique. Il existe deux types principaux de symbiose mycorhizienne. Dans le cas des endomycorhizes (ou mycorhizes arbusculaires), les hyphes du champignon pénètrent à l'intérieur des cellules des racines de la plante et forment des arbuscules dans le cytoplasme où ont lieu les échanges de nutriments. Cette symbiose est prédominante et est formée par 70 à 90 % des plantes végétales terrestres avec le phylum symbiotique obligatoire *Glomeromycota* (Shi et al., 2023). Dans le cas des ectomycorhizes, les hyphes du champignon entourent les racines de la plante sans pénétrer à l'intérieur des cellules et forment un manchon autour des racines, le réseau de Hartig, au niveau duquel ont lieu les échanges de nutriments. Cette symbiose concerne environ 2 % des plantes terrestres mais joue un rôle écologique majeur puisqu'elle concerne plus de la moitié des espèces d'arbres, et quasiment la totalité des espèces d'arbres en climat tempéré ou boréal. Enfin, les champignons saprotrophes sont des organismes fongiques qui se nourrissent en décomposant et en absorbant la matière organique morte ou en décomposition dans son environnement. Ils contribuent à l'essentiel du recyclage des nutriments minéraux de la litière et à jusqu'à 90 % de la respiration hétérotrophe totale dans les écosystèmes forestiers (Cooke et Rayner, 1984).

La grande majorité des espèces fongiques nommées et décrites à ce jour sont susceptibles d'être présentes dans le sol à un moment ou un autre de leur cycle de vie (Bridge & Spooner, 2001). Elles jouent un rôle central dans de nombreux processus, influençant la fertilité du sol, la santé des plantes et leur nutrition, ou en décomposant et recyclant des minéraux et de la matière organique. Elles sont très diversifiées aussi bien structurellement que fonctionnellement (Finlay & Thorn, 2019).

Comprendre comment les communautés fongiques fonctionnent est crucial pour la biologie du sol. Cela peut aider par exemple à identifier des leviers pour améliorer le stockage du carbone dans les sols, ce qui serait un outil efficace pour atténuer le changement climatique. Cette compréhension implique l'utilisation d'approches fonctionnelles, notamment au travers d'études métatranscriptomiques, dans lesquelles la capacité à identifier les fonctions exprimées par chaque guilde écologique est cruciale (Auer et al, 2023).

Les études sur l'écologie fongique se sont, jusqu'à très récemment, souvent focalisées sur des guildes trophiques particulières. Il est par exemple courant que les études d'écologie forestière se concentrent

uniquement sur les champignons mycorhiziens dans un échantillon de sol ou sur les décomposeurs dans le bois. Cependant, comme les guildes interagissent souvent entre elles (Clemmensen et al., 2015), et ont souvent des réponses contrastées aux mêmes gradients environnementaux (Šťursová et al., 2014), les analyses globales des communautés fongiques totales passent à côté de tendances écologiques importantes qui s'annulent ou sont masquées lorsqu'elles sont considérées dans leur ensemble. Certains articles récents basés sur des analyses à haut débit ont reconnu ce fait et ont commencé à analyser leurs pools de séquences par guilde, reconnaissant les saprotrophes, les champignons ectomycorhiziens, les agents pathogènes des plantes, les agents pathogènes des animaux, les mycoparasites, les saprotrophes du bois, les champignons lichénisés, et même les levures et les endophytes septés sombres (Nguyen et al., 2016).

Des travaux précédents du laboratoire INRAE Champenoux à Nancy (Auer et al, 2023) se sont penchés sur la compétition pour les ressources organiques pour les guildes saprotrophes, pathogènes et symbiotes mycorhiziens dans des sols forestiers boréaux, tempérés et méditerranéens. Aucune étude n'avait jusqu'à présent comparée l'expression génique actuelle de ces guildes dans les différents sols forestiers, les champignons jouant un rôle majeur dans les deux principaux processus de maintien de la vie des arbres dans ces écosystèmes : directement par absorption des nutriments et croissance des arbres, et indirectement par décomposition. Il a été mis en lumière que les champignons ectomycorhiziens et saprotrophes sont en compétition pour l'azote dans la matière organique du sol, suggérant que leurs interactions à certains endroits peuvent ralentir le cycle du carbone, cycle vital pour maintenir l'équilibre des concentrations de dioxyde de carbone dans l'atmosphère.

Ces résultats, établissant des liens entre expression génique et traits fonctionnels pour les champignons (des études comme celle-ci n'avaient été menées que pour les bactéries), ont été obtenu par une approche de métatranscriptomique, des opportunités d'étudier les diversités taxonomiques et fonctionnelles dans les sols ont été possible via des approches d'annotation sur référence. Pour cela il a été fait usage d'un pipeline d'annotation de l'ARNm combiné à la base de données MycoCosm, issue du 1000 Fungal Genome. L'approche d'annotation choisie repose sur la recherche d'homologies entre les transcrits partiels, reconstruits par assemblage des lectures de séquençage, et les gènes répertoriés dans cette base de données.

Cette base de données comprend en 2024 plus de 2500 génomes annotés et de plus, de façon assez homogène. Cependant, bien que cette ressource représente un atout précieux, elle est à relativiser au regard des 3,8 millions d'espèces de champignons présentes sur Terre (Nilsson et al., 2019). Il est ainsi nécessaire de pouvoir mesurer la fiabilité des prédictions d'annotations taxonomiques, trophiques et fonctionnelles que l'on peut obtenir pour des échantillons comportant des centaines d'espèces aux génomes majoritairement inconnus.

L'objectif de ce projet est donc le développement d'une approche comparative entre annotations attendues et prédites, en retirant un à un les génomes de MycoCosm avant de les interroger contre le reste de la base de données, et ainsi mesurer la fiabilité de l'approche d'annotation sur référence.

2. Matériels et méthodes :

a. Constitution et nettoyage de la base de données MycoCosm 2024

2526 génomes fongiques ont été téléchargés à partir du portail web de MycoCosm et stockés sur le serveur INRAE Champenoux le 12/01/2024, en autant de répertoires contenant les fichiers fastas de séquences codantes (CDS), de séquences d'acides aminés (AA), et les tableaux d'annotations (KOG et Signal-peptide). L'approche générale a consisté en la réalisation d'un Diamond Blastx impliquant la séquence codante (CDS) de chaque génome contre une base de données de séquences AA de tous les génomes hormis le sien, afin de correspondre à la méthode utilisée dans le laboratoire sur des échantillons de métatranscriptomique, pour lesquels après assemblage et filtres, les séquences nucléotidiques sont interrogées contre une base de données protéique.

La réalisation des Diamonds Blastx nécessitait donc un fichier AA et un fichier CDS par répertoires de génomes. Sur les 2526 génomes téléchargés, 27 répertoires de génomes ne contenaient pas de paires CDS/AA valides et ont été exclus de l'analyse. 16 répertoires contenaient plusieurs fichiers CDS et/ou AA, dans ces cas, un nettoyage manuel a permis de ne conserver qu'une seule paire CDS/AA (la plus récente ou le génome haploïde). 2483 répertoires possédaient le bon nombre de fichiers AA/CDS.

Finalement, 4998 fichiers d'extensions «.fasta.gz », soit 2 pour 2499 répertoires de génomes ont ainsi été conservés. Courant mois de janvier, 20 nouveaux génomes ont été ajoutés à la base de données MycoCosm. Ils ont été ajoutés à la liste de génomes soumis à l'analyse après formatage de leurs noms de fichiers, faisant passer le nombre de fichiers contenant les CDS et les AA a 5038, pour 2519 répertoires de génomes.

b. Réalisation des 2519 Diamonds BlastX

Pour chaque génome, une base de référence et l'index Diamond correspondant sont construits à partir des 2518 fichiers AA des autres génomes. Le fichier CDS de ce génome est ensuite utilisé pour un Blastx avec Diamond (version 2.0.15, paramètres par défaut excepté le seuil d'e-value, à 10^{-10}) à l'identique de protocole suivi précédemment dans le laboratoire (Auer et al., 2023) contre l'index obtenu. Une fois les sorties Diamonds obtenues (meilleur hit pour chaque gène du génome et métriques associées), la base et l'index de référence (resp. 10 et 18 Go) sont supprimés et le processus complet est répété pour chaque génome. L'opération sur les 2519 génomes a été divisée en 3 lots et

parallélisée sur 3 * 40 CPU sur le cluster de calcul du laboratoire, pour un temps d'exécution d'environ 20 jours de calcul (soit environ 60 000 heures CPU).

c. Annotations fonctionnelles KOG et SignalP

En partant des 2546 répertoires et après un processus de contrôle des fichiers et d'élimination des doublons de versions, 2521 fichiers KOG et 2518 fichiers SigP ont été isolés de la base de données MycoCosm.

Les annotations KOG et SigP sont ensuite assignées aux fichiers sorties Diamonds. Les fichiers sortie Diamonds se composent de l'identification de la 'query', qui contient l'abréviation de la base de données ('jgi' pour 'Joint Genome Institute'), le nom du génome d'origine du gène, le transcriptID du gène au sein de ce génome, et enfin une description. La query suit cette structure type : jgi|genome|transcriptID|descr. Séparée par une tabulation, la query est associée à son meilleur 'hit' contre la base de données, qui respecte la même structure type : jgi|genome|protID|descr, à la différence près qu'elle fait référence à un identifiant de protéine et non de gène (leur numérotation étant indépendante). Ces identifiants se retrouvent également dans les tables d'annotation, associés aux annotations KOG et SigP. Par un système de dictionnaires, les identifiants de transcrit, de protéines et les annotations KOG et SigP sont croisées pour extraire pour chaque gène d'un génome ses annotations attendues (à partir de l'identification de la query) et prédites (à partir de l'identification du meilleur hit). L'opération est appliquée aux sorties Diamond Blastx des 2519 génomes.

d. Calculs des pourcentages de bonnes prédictions

Les annotations taxonomiques et trophiques des génomes sont obtenues via la base de données Fungaltraits 2024 et sont assignées à chaque queries et hits des fichiers sortie Diamonds. La plus proche référence taxonomique 'closest-ref' de chaque génome de la base de données est récupérée en cherchant pour chaque génome les éléments de sa taxonomie partagés avec d'autres génomes. Cette information indique ainsi pour chaque génome le niveau taxonomique commun avec la référence la plus proche (le genre si d'autres génomes du même genre sont présents dans la base, ou la famille, etc. jusqu'au phylum). Cette information est essentielle à la suite de l'analyse, puisque

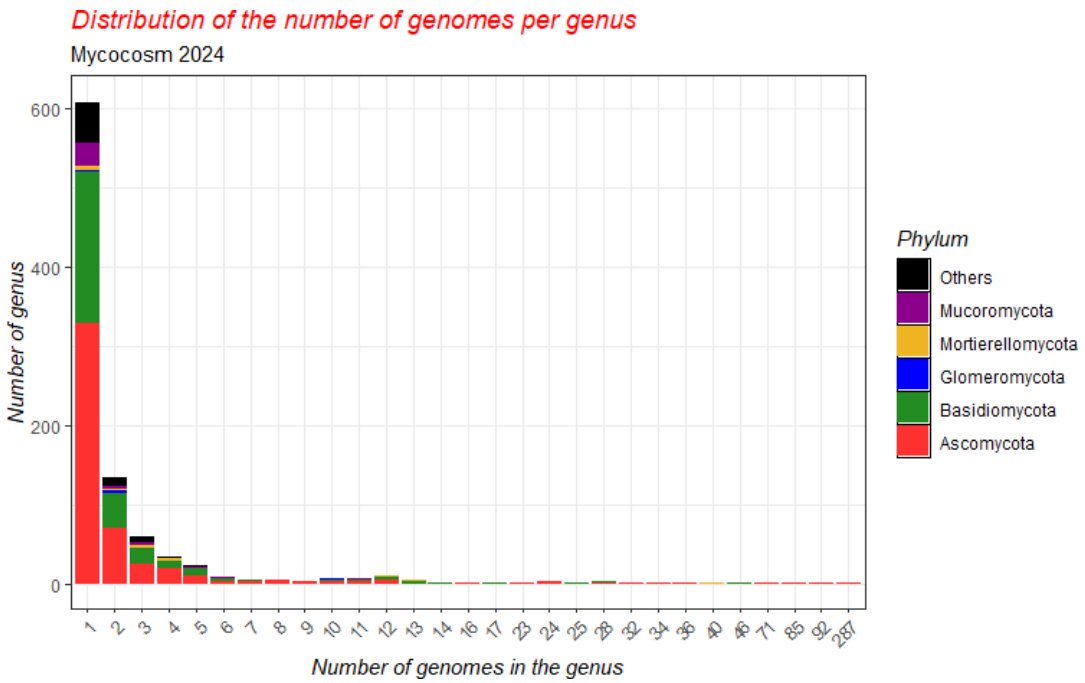

Figure 2 : Représentation barplot de la distribution du nombre de genres en fonction du nombre de génomes par genres dans la base de données et en fonction des différents phylas d'appartenance (représentés par différentes couleurs)

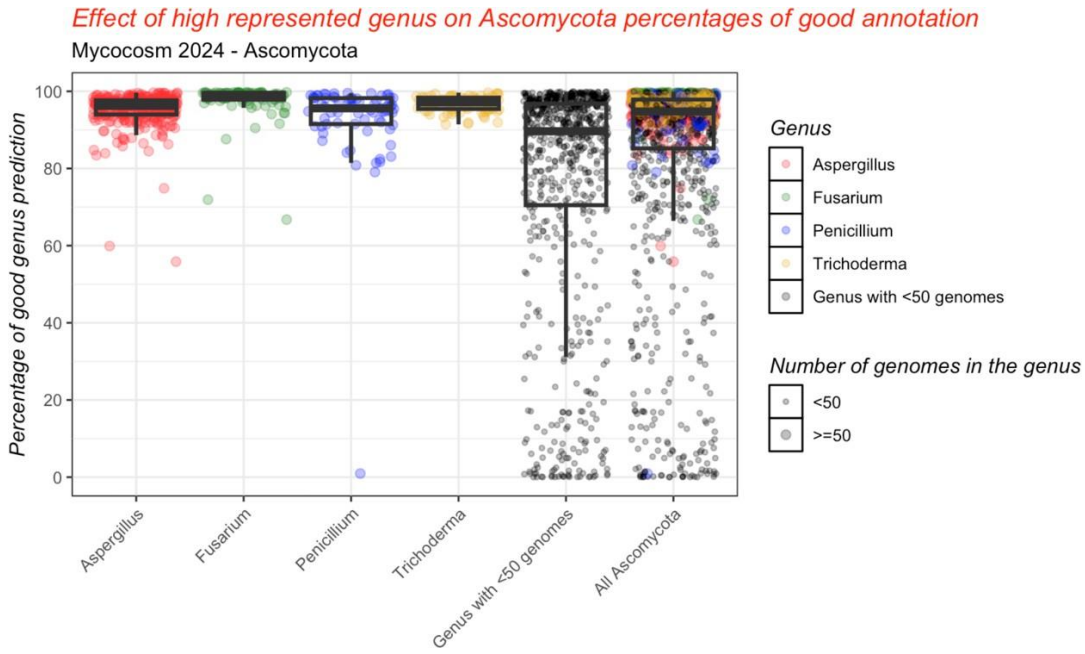

Figure 3 : Représentation boxplots de l'effet des genres sur-représentés du phylum *Ascomycota* sur les pourcentages de bonnes prédictions au genre. Les genres *Aspergillus, Fusarium, Penicillium* et *Trichoderma* possèdent plus de 50 références et leurs génomes sont représentés dans des couleurs distinctes. Les génomes des genres du phylum *Ascomycota* possédant moins de 50 références sont représentés en noir. Les pourcentages de bonnes prédictions au genre en incluant tous les génomes du phylum *Ascomycota* sont ensuite calculés (boxplot de droite), les génomes sont discriminés en fonction du nombre de références qu'ils contiennent par la taille.

mécaniquement, un génome qui n'a pas de « closest_ref » à un niveau taxonomique donné ne peut pas avoir de bonnes prédictions à ce niveau ainsi qu'aux niveaux inférieurs.

Une information de « Guilde » a été générée pour chaque génome. A partir des informations de style de vie primaire et secondaire récupérées dans Fungaltraits (32 modalités), une guilde est assignée à chaque génome de la base de données parmi Saprotrophe (regroupant 6 modalités), Ectomycorrhizien (1 modalité), Pathotrophe (2 modalités) ou Autres (tout le reste).

A partir des fichiers de sortie Diamonds, des informations taxonomiques et trophiques et des annotations KOG et SigP, une table contenant les pourcentages de gènes sans prédiction d'annotation (%NoHit), de bonnes prédictions aux différents niveaux taxonomiques, à la guilde et aux fonctions pour tous les génomes Mycocosm 2024 est générée via un script R "data.table", une librairie spécifiquement développée pour traiter les très gros jeux de données et nécessaire ici.

A partir de ces données, de nombreuses analyses exploratoires sont réalisées en langage R et à l'aide de la librairie ggplot afin de visualiser les résultats obtenus. Un test Wilcoxon avec correction FDR a été appliqué pour certains résultats pertinents. Toutes les analyses sont réalisées dans un Markdown organisé en menus déroulant permettant leur traçabilité.

3. Résultats :

a. Taxonomie :

16 phylas ont été recensés à partir des génomes analysés. Parmi eux, certains contenaient peu de génomes comparativement aux autres, et ils ont donc été rassemblés dans une catégorie 'Others' tandis que les phylums *Ascomycota*, *Basidiomycota*, *Glomeromycota*, *Mortierellomycota* et *Mucoromycota* ont été conservés.

Les 2519 génomes de MycoCosm ne sont pas répartis de façon uniforme dans l'arbre des champignons, et en particulier, certains genres d'intérêt sont largement sur-représentés, avec plusieurs dizaines de génomes, jusqu'à 287 pour *Aspergillus* du phylum *Ascomycota* (Figure 2). Sur les 889 genres que contient la base de données, 601 ne sont par ailleurs représentés que par un seul génome dans MycoCosm, signifiant que ces génomes ne possèdent pas de plus proches références taxonomiques au genre. Cette inéquité dans la distribution des génomes pourrait être source de biais dans les analyses par la suite. Afin d'évaluer l'effet de ce biais, une mesure des pourcentages de bonnes prédictions au genre pour les genres du phylum *Ascomycota*, en fonction du nombre de référence qu'ils contiennent, a été réalisée (figure 3). 4 genres Ascomycètes contiennent plus de 50 génomes : *Aspergillus, Fusarium, Penicillium* et *Trichoderma, et* la médiane de leur pourcentage de gènes bien annotés au genre est de 98%, 94% et 96% respectivement. La comparaison des

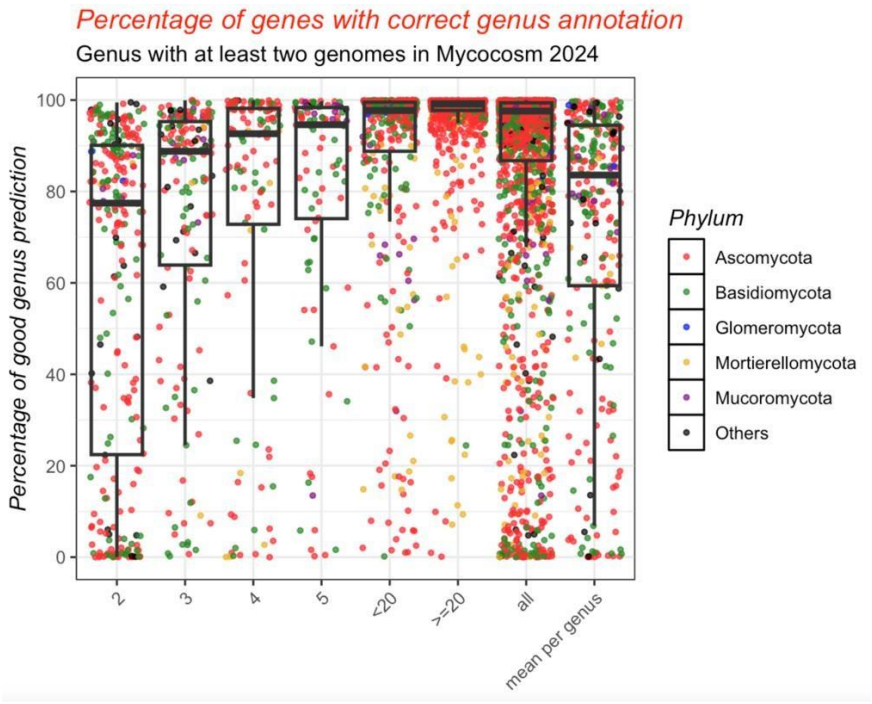

Figure 4 : Représentations boxplots des pourcentages de bonnes prédictions au genre en fonction du nombre de références que contiennent les genres. Les différents phylas auxquels appartiennent les genres sont représentés par différentes couleurs. Le boxplot "all" représente les pourcentages de bonnes prédictions au genre pour tous les génomes de la base de données tandis que le boxplot "mean per genus" représente les pourcentages de bonnes prédictions au genre pour toutes les moyennes de pourcentages de bonnes prédictions au genre par genres.

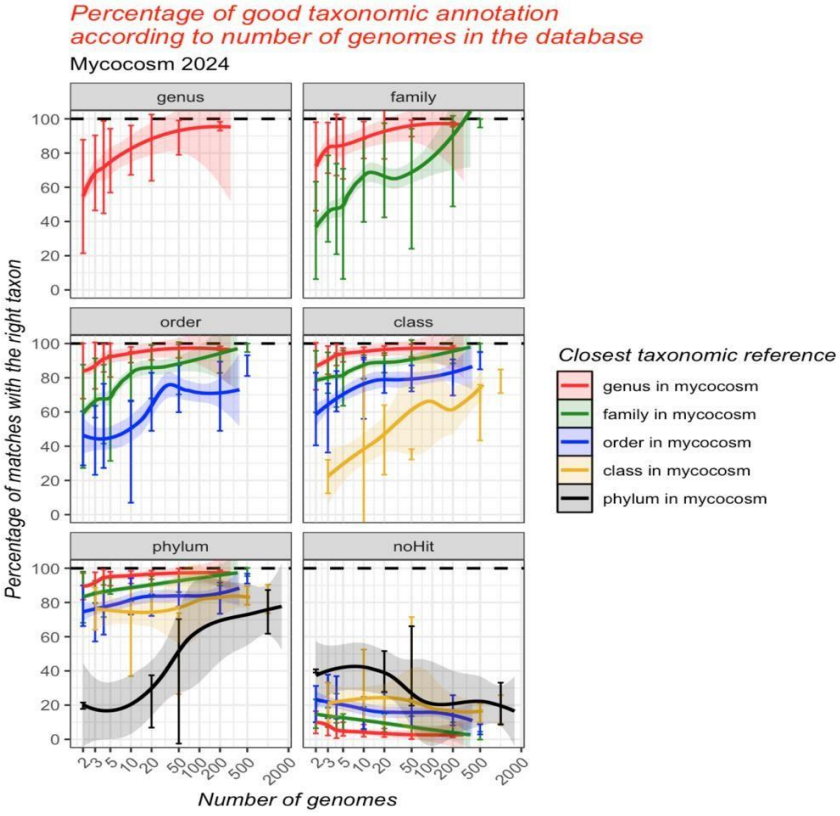

Figure 5 : Représentations par graphiques courbes des pourcentages de bonnes prédictions à tous les niveaux de comparaison taxonomiques (du genre au phylum) en fonction de la plus proche référence taxonomique des différents taxons de la base de données (genre lorsque le genre est présent dans MycoCosm, famille lorsque le genre n'est pas présent mais la famille est présente, etc...) et en fonction du nombre de référence par taxons. Les différents taxons sont représentés par des couleurs différentes.

distributions de pourcentages de bonnes prédictions au genre pour les genres *Ascomycota* comportant moins de 50 références d'une part, et pour tous les genres *Ascomycota* d'autre part, montre l'influence de la présence de ces 4 genres sur-représentés. Les pourcentages de bonnes prédictions au genre pour le phylum *Ascomycota* passent d'une médiane de 89% lorsque l'on considère que les genres possédant moins de 50 références à 94% lorsque l'on considère tous les genres du phylum. Du fait de cette augmentation et des fortes disparités en nombre de génomes par genre, l'analyse des distributions au niveau génome présente un fort risque de biais : *Aspergillus,* représenté 287 fois avec un très bon pourcentage, "tire" la distribution vers le haut, ce qui peut fausser l'appréciation de la qualité d'annotation. Pour mesurer cet effet, on peut étudier l'évolution du pourcentage de bonnes prédictions taxonomiques en fonction du nombre de génomes de référence. Ainsi, pour les 288 genres pour lesquels au moins un autre génome du même genre est disponible, on peut suivre le pourcentage de gènes avec une bonne prédiction de genre, en fonction du nombre de génomes de référence disponibles dans le genre (figure 4). On observe une augmentation du pourcentage de bonnes prédictions au genre lorsque le nombre de génomes dans le genre augmente, d'une médiane de 78% pour les génomes avec une seule autre référence (2 génomes dans le genre) à 99% pour les génomes avec plus de 20 références au genre. La distribution des pourcentages de bonnes prédictions au genre pour tous les génomes de la base de données présente une très longue queue de distribution, mais du fait des genres contenant des centaines de génomes, plus de la moitié des génomes ont un pourcentage de bonne annotation de plus de 98%. Suivre la moyenne par genre permet de gommer cet effet : on observe une forte diminution, mais qui correspond mieux à la réalité biologique des données (même si elle masque les disparités au sein d'un genre). Les phylum *Ascomycota* et *Basidiomycota* montrent une grande distribution du nombre de références par genre. Les genres du phylum *Mortierellomycota* et *Mucoromycota* présentent en majorité de nombreuses références par genres et les genres du phylum *Glomeromycota* et de la catégorie "Others" possèdent 2 à 10 références par genre seulement.

Afin de savoir si cette tendance est spécifique au genre ou conservée pour tous les niveaux taxonomiques, l'augmentation du pourcentage de bonnes prédictions à tous les niveaux taxonomiques en fonction de l'augmentation du nombre de référence dans chaque taxon, pour tous les génomes de MycoCosm 2024 a été mesurée (figure 5). La tendance de l'augmentation du pourcentage de bonnes prédictions en fonction du nombre de références est conservée à tous les niveaux de comparaison taxonomiques, avec des décalages en fonction de la distance à la référence la plus proche : Les pourcentages de bonnes prédictions au taxon sont toujours meilleurs lorsque la plus proche référence taxonomique est le genre, et ils s'améliorent selon que le niveau de comparaison taxonomique est large. Un génome ne possédant pas de plus proche référence taxonomique à un taxon donné ne peut pas avoir de bonnes prédictions à ce niveau ainsi qu'aux

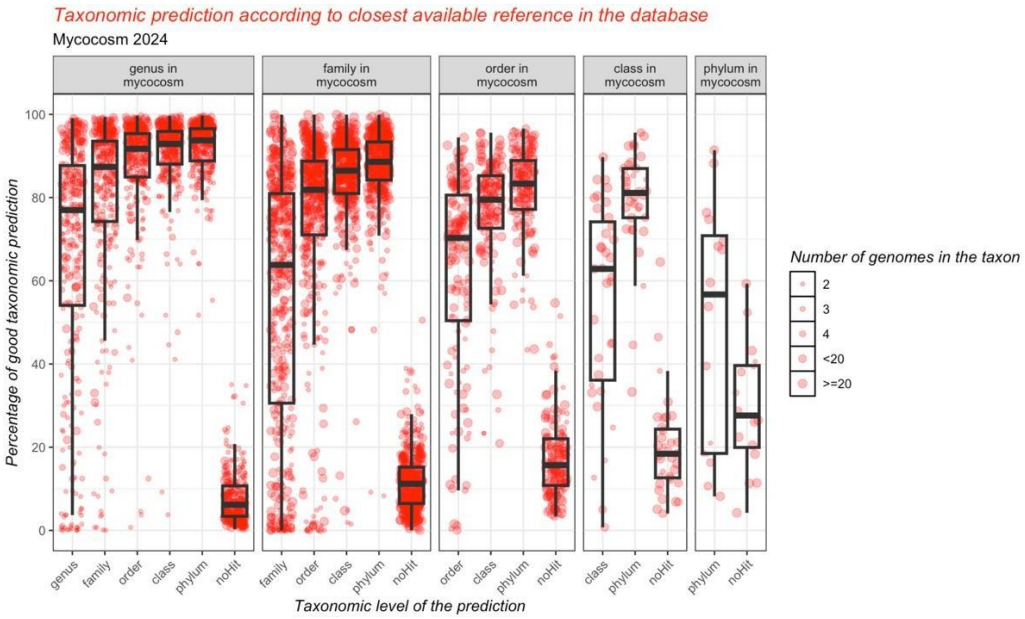

Figure 6 : Représentation boxplots du pourcentage de bonnes prédictions à tous les niveaux taxonomiques, pour tous les taxons de la base de données, en fonction du nombre de références qu'ils contiennent et du niveau de plus proche référence taxonomique. Les taxons sont représentés par des tailles différentes selon le nombre de référence qu'ils contiennent et la plus proche référence taxonomique des taxons est représentée par des couleurs différentes.

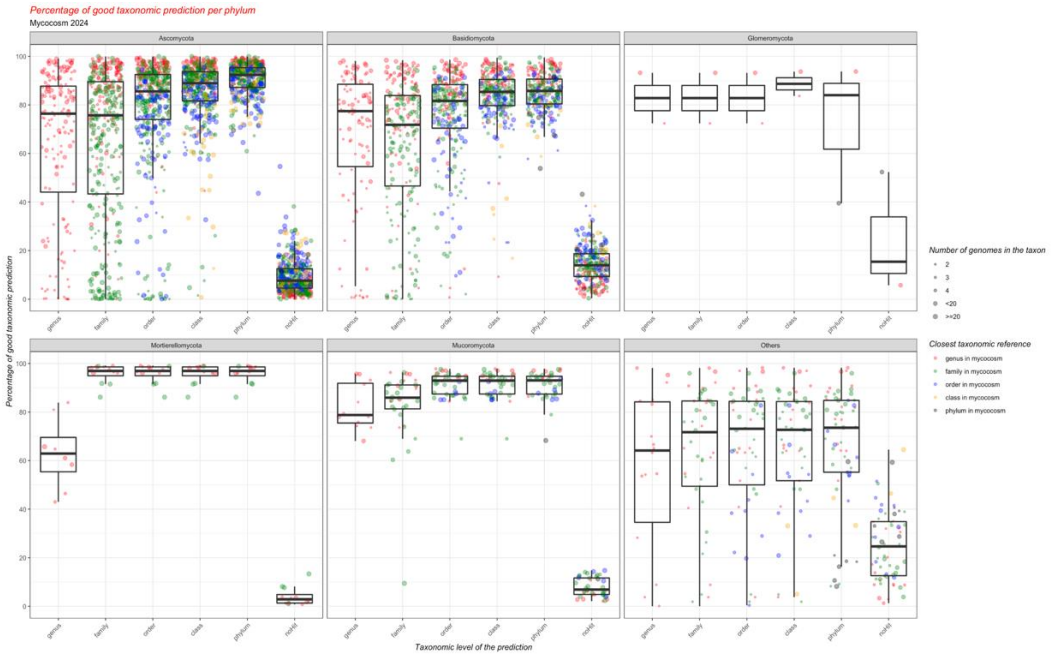

Figure 7 : Représentation boxplots du pourcentage de bonnes prédictions à tous les niveaux taxonomiques, pour tous les taxons de la base de données, en fonction du nombre de références qu'ils contiennent et du niveau de plus proche référence taxonomique, par phylas. Les taxons sont représentés par des tailles différentes selon le nombre de référencequ'ils contiennent et la plus proche référence taxonomique des taxons est représentée par des couleurs différentes.

niveaux inférieurs. Les pourcentages de Faux Négatifs (noHit) augmentent lorsque la plus proche référence taxonomique des génomes s'éloigne du genre, de 3 à 10% pour le genre (en fonction du nombre de génomes disponibles) à 20 à 40% pour le phylum.

L'effet distordant des genres très représentés étant établi, il a été décidé pour la suite des analyses d'observer des moyennes de pourcentages par genre, ce qui permet de donner à chaque genre le même poids sur les distributions, et de compenser les disparités de nombre de génomes de référence par genre. L'information du nombre de génomes par genre (mais également par niveau taxonomique, de la famille au phylum) a néanmoins été conservée et étudiée.

L'effet de la distance à la plus proche référence sur la distribution du pourcentage par genre de bonnes prédictions à chaque niveau taxonomique peut ainsi être étudié (figure 6). Une importante queue de distribution correspondant aux taxons les moins bien prédits est observée pour le boxplot de pourcentages de bonnes prédictions au genre lorsque la plus proche référence est le genre ainsi que pour le boxplot de pourcentages de bonnes prédictions à la famille lorsque la plus proche référence est la famille. Certains des genres et familles mal prédits contiennent pourtant de nombreuses références dans la base de données. Les pourcentages de bonnes prédictions augmentent quand le niveau taxonomique prédit est plus large, mais plafonnent du fait du pourcentage de gènes sans annotation prédite, également représenté (noHit), qui augmente fortement avec la distance à la référence.

Le même type d'analyse, mais en séparant les genres en fonction de leur phylum d'appartenance a été réalisé afin d'étudier si les différents phylas présentent les mêmes comportements (figure 7). Les phylas *Ascomycota* et *Basidiomycota* formant le groupe des *Dicaria* contiennent le plus de références pour tous les taxons. Ils obtiennent de bons résultats de prédiction au genre (la médiane est à 77%) avec des pourcentages de Faux négatifs (%noHit) assez faibles (médiane à 6% pour *Ascomycota*, 10% pour *Basidiomycota*). On observe cependant une queue de distribution assez importante correspondante à des genres mal prédits au niveau du boxplot de comparaison au genre chez les *Ascomycota*, certains de ces genres comportant pourtant de nombreuses références. Ces mauvais pourcentages de prédiction au taxon tendent à disparaître pour les génomes dont la plus proche référence taxonomique est le genre lorsque le niveau de comparaison est l'ordre, pour les phylas *Ascomycota* et *Basidiomycota*. Les pourcentages de bonnes prédictions à tous les niveaux sont bons pour le phylum *Glomeromycota* (les médianes sont au-dessus de 80% à tous les niveaux taxonomiques). Le phylum *Mortierellomycota* montre de faibles pourcentages de bonnes prédictions au genre avec une médiane à 64%, malgré de nombreuses références contenues dans certains genres mal prédits (le genre *Mortierella* possède 40 références dans la base de données mais obtient un pourcentage de bonnes prédictions de 65,74% seulement). Les pourcentages de

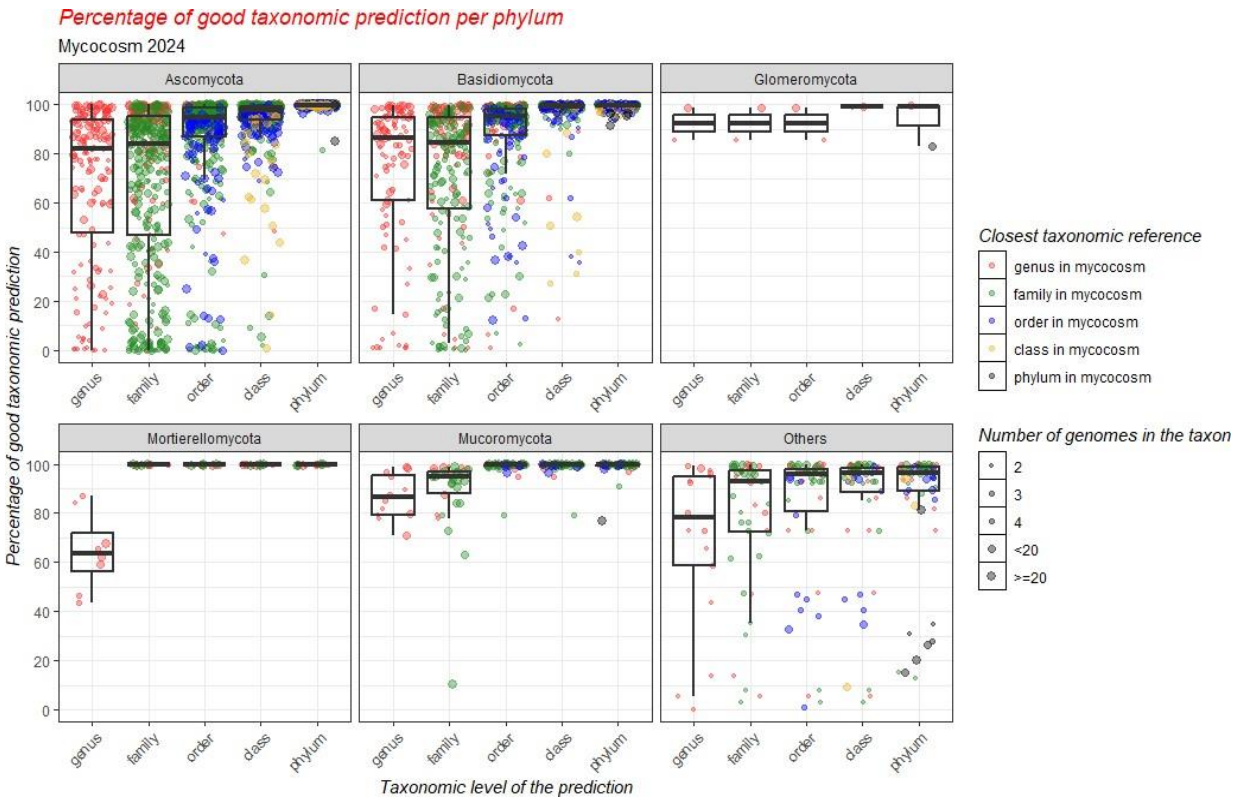

Figure 8 : Représentation boxplots du pourcentage de bonnes prédictions à tous les niveaux taxonomiques, pour tous les taxons de la base de données, en fonction du nombre de références qu'ils contiennent et du niveau de plus proche référence taxonomique, par phylas. Les gènes n'ayant pas obtenus de meilleurs hits ("noHits") ont été exclus. Les taxons sont représentés par des tailles différentes selon le nombre de référence qu'ils contiennent et la plus proche référence taxonomique des taxons est représentée par des couleurs différentes.

	Genus	Percentages_means_genus	GenusPred	FamilyPred	OrderPred	ClassPred	PhylumPred	sum_genus_pred_per_genus	Family	Order	Class	Phylum
1	Anomoporia	41.758089	Anomoporia	Amylocorticiaceae	Amylocorticiales	Agaricomycetes	Basidiomycota	15852	Amylocorticiaceae	Amylocorticiales	Agaricomycetes	Basidiomycota
2	Anomoporia	41.758089	Anomoloma	Amylocorticiaceae	Amylocorticiales	Agaricomycetes	Basidiomycota	9809	Amylocorticiaceae	Amylocorticiales	Agaricomycetes	Basidiomycota
3	Cadophora	28.798918	Cadophora	Ploettnerulaceae	Helotiales	Leotiomycetes	Ascomycota	20148	Ploettnerulaceae	Helotiales	Leotiomycetes	Ascomycota
4	Cadophora	28.798918	Leptodontidium	Leptodontidiaceae	Helotiales	Leotiomycetes	Ascomycota	15565	Ploettnerulaceae	Helotiales	Leotiomycetes	Ascomycota
5	Capronia	33.506847	Capronia	Herpotrichiellaceae	Chaetothyriales	Eurotiomycetes	Ascomycota	12917	Herpotrichiellaceae	Chaetothyriales	Eurotiomycetes	Ascomycota
6	Capronia	33.506847	Cladophialophora	Herpotrichiellaceae	Chaetothyriales	Eurotiomycetes	Ascomycota	11390	Herpotrichiellaceae	Chaetothyriales	Eurotiomycetes	Ascomycota
7	Cercophora	26.405399	Podospora	Podosporaceae	Sordariales	Sordariomycetes	Ascomycota	35107	Lasiosphaeriaceae	Sordariales	Sordariomycetes	Ascomycota
8	Cercophora	26.405399	Cercophora	Lasiosphaeriaceae	Sordariales	Sordariomycetes	Ascomycota	17368	Lasiosphaeriaceae	Sordariales	Sordariomycetes	Ascomycota
9	Diaporthe	47.366978	Diaporthe	Diaporthaceae	Diaporthales	Sordariomycetes	Ascomycota	18800	Diaporthaceae	Diaporthales	Sordariomycetes	Ascomycota
10	Diaporthe	47.366978	Unclassified_DiaPMI573_1	Diaporthaceae	Diaporthales	Sordariomycetes	Ascomycota	14600	Diaporthaceae	Diaporthales	Sordariomycetes	Ascomycota
11	Entomortierella	43.616218	Entomortierella	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota	19712	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota
12	Entomortierella	43.616218	Mortierella	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota	13124	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota
13	Erysiphe	47.561771	Erysiphe	Erysiphaceae	Erysiphales	Leotiomycetes	Ascomycota	9444	Erysiphaceae	Erysiphales	Leotiomycetes	Ascomycota
14	Erysiphe	47.561771	Oidium	Erysiphaceae	Erysiphales	Leotiomycetes	Ascomycota	7256	Erysiphaceae	Erysiphales	Leotiomycetes	Ascomycota
15	Haplosporangium	46.652167	Haplosporangium	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota	18121	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota
16	Haplosporangium	46.652167	Mortierella	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota	10300	Mortierellaceae	Mortierellales	Mortierellomycetes	Mortierellomycota
17	Hortaea	49.325587	Hortaea	Teratosphaeriaceae	Capnodiales	Dothideomycetes	Ascomycota	17460	Teratosphaeriaceae	Capnodiales	Dothideomycetes	Ascomycota
18	Hortaea	49.325587	Extremus	Extremaceae	Capnodiales	Dothideomycetes	Ascomycota	3176	Teratosphaeriaceae	Capnodiales	Dothideomycetes	Ascomycota
19	Meliniomyces	17.596940	Hyaloscypha	Hyaloscyphaceae	Helotiales	Leotiomycetes	Ascomycota	21051	Hyaloscyphaceae	Helotiales	Leotiomycetes	Ascomycota
20	Meliniomyces	17.596940	Chalara	Pezizellaceae	Helotiales	Leotiomycetes	Ascomycota	14203	Hyaloscyphaceae	Helotiales	Leotiomycetes	Ascomycota
21	Ophiocordyceps	43.151405	Ophiocordyceps	Ophiocordycipitaceae	Hypocreales	Sordariomycetes	Ascomycota	12750	Ophiocordycipitaceae	Hypocreales	Sordariomycetes	Ascomycota
22	Ophiocordyceps	43.151405	Tolypocladium	Ophiocordycipitaceae	Hypocreales	Sordariomycetes	Ascomycota	7023	Ophiocordycipitaceae	Hypocreales	Sordariomycetes	Ascomycota
23	Phellinus	14.553702	Porodaedalea	Hymenochaetaceae	Hymenochaetales	Agaricomycetes	Basidiomycota	16386	Hymenochaetaceae	Hymenochaetales	Agaricomycetes	Basidiomycota
24	Phellinus	14.553702	Onnia	Hymenochaetaceae	Hymenochaetales	Agaricomycetes	Basidiomycota	8029	Hymenochaetaceae	Hymenochaetales	Agaricomycetes	Basidiomycota
25	Pichia	15.587848	Issatchenkia	Saccharomycetaceae	Saccharomycetales	Saccharomycetes	Ascomycota	11104	Pichiaceae	Saccharomycetales	Saccharomycetes	Ascomycota
26	Pichia	15.587848	Pichia	Pichiaceae	Saccharomycetales	Saccharomycetes	Ascomycota	3084	Pichiaceae	Saccharomycetales	Saccharomycetes	Ascomycota

Tableau 1 : Tableau récapitulant les pourcentages de bonnes prédictions au genre pour les genres contenant au moins 3 références les moins bien prédits (<50%), leur taxonomie ainsi que la taxonomie de leurs meilleurs et deuxièmes meilleurs hits (les nombres de hits au genre sont obtenus à partir de la colonne "sum_genus_pred_per_genus").

bonnes prédictions grimpent rapidement pour ce phylum dès que l'on observe des prédictions à la famille ou à un taxon plus large avec des médianes supérieures à 95%). Le phylum *Mucoromycota* obtient de bonnes prédictions pour le genre et la famille (médianes à 78% et 82% respectivement) et de très bons pourcentages de bonnes prédictions pour les autres niveaux de comparaison (médianes supérieures à 92%). Concernant les phylas regroupés dans la catégorie "Others" les pourcentages de bonnes prédictions à tous les niveaux taxonomiques sont moins bons (les moyennes varient de 65% à 74%) et les pourcentages de faux négatifs sont plus élevés que pour le reste des phylas (médiane à 20 %). On observe cependant une sur-représentation des taxas ayant de nombreuses références parmi les taxons les mieux prédits, à tous les niveaux de comparaison taxonomiques.

En conditions réelles avec des échantillons de terrain, les transcrits sans hit sur Mycocosm peuvent être des transcrits non fongiques ou des artefacts d'assemblage. Ils sont donc nécessairement éliminés, et le pourcentage de faux négatifs n'est pas forcément pertinent à intégrer. Il a été décidé pour la suite des analyses d'observer des pourcentages de bonnes prédictions sur la base d'annotations prédites uniquement (les noHit ont été exclus), ce qui permet d'étudier plus spécifiquement les niveaux de bonnes et de mauvaises prédictions, en écartant la question des gènes non prédits.

Ainsi, les pourcentages de bonnes prédictions à tous les niveaux, pour tous les taxons et pour tous les phylums ont été recalculés à partir des gènes ayant reçus une annotation uniquement (figure 8). On observe pour le phylum *Ascomycota*, une augmentation du pourcentage de bonnes prédictions au genre (la médiane passe de 78% à 81%). Les pourcentages de bonnes prédictions au phylum atteignent plus de 99% pour ce phylum. Les pourcentages de bonnes prédictions au genre pour le phylum *Basidiomycota* passent de 78% à 86%. Certains phylas tel *Mortierollomycota* ont ainsi vu leurs pourcentages de bonnes prédictions taxonomiques très peu évoluer.

Comme mentionné précédemment, des queues de distribution assez longues persistent, même pour des genres avec beaucoup de références proches. Les genres avec les pourcentages de bonnes prédictions les plus faibles (inférieurs à 50 %) et au moins 3 génomes de référence ont été récupérés pour une recherche des causes, en analysant notamment leurs mauvaises prédictions majoritaires (tableau 1). Ainsi les génomes appartenant au genre *Anomosporia* obtiennent un match Diamonds blast de leurs gènes avec des gènes de génomes appartenant à ce même genre en premier et en second lieu. Ils obtiennent aussi beaucoup de "mismatchs" au genre car leur pourcentage moyen de bonnes prédictions au genre est de 41,75 % seulement. Les gènes des génomes appartenant au genre *Phellinus* ne matchent pas préférentiellement avec des gènes de génomes du même genre (*Porodaedalea* et *Onnia* en premier et second lieu, 16386 et 8029 fois respectivement) mais avec des gènes de génomes de la même famille (*Hymenochaetaceae*).

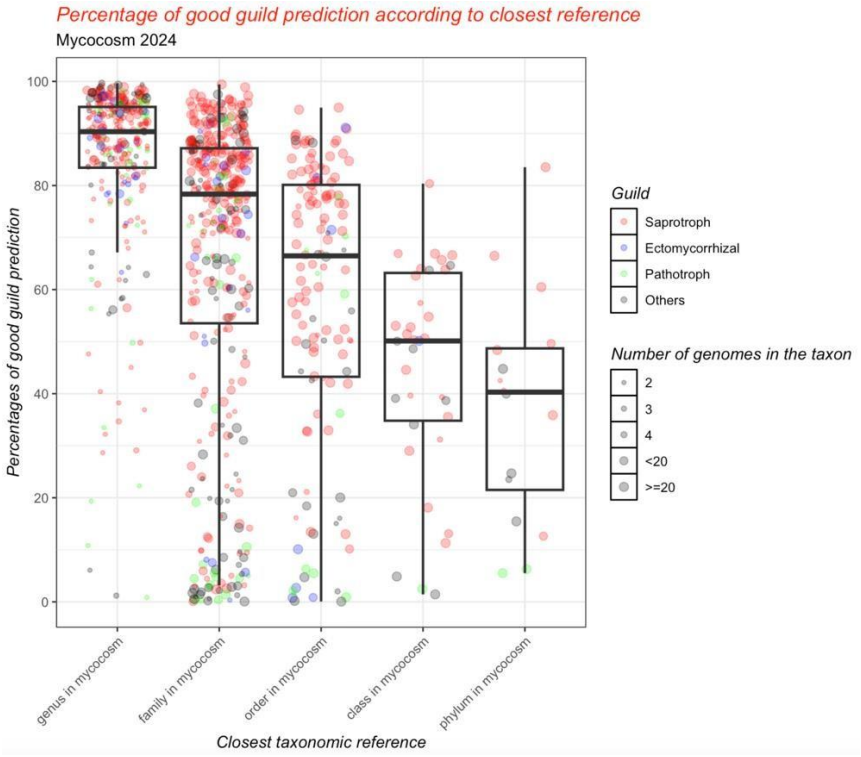

Figure 9 : Représentation boxplots des pourcentages de bonnes prédictions à la guilde en fonction du plus proche niveau de référence taxonomique, du nombre de référence par taxons et de la guilde. Les taxons sont représentés par des tailles différentes selon le nombre de références qu'ils contiennent et la guilde est représentée par des couleurs différentes.

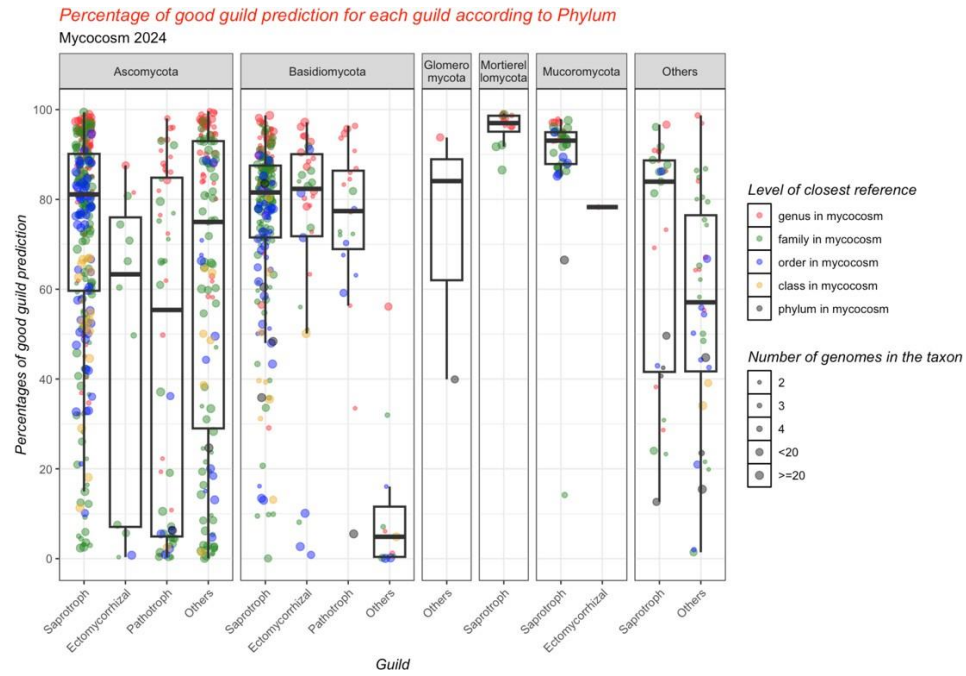

Figure 10 : Représentation boxplots des pourcentages de bonnes prédictions à la guilde en fonction du plus proche niveau de référence taxonomique, du nombre de référence par taxons, de la guilde et par phylas. Les taxons sont représentés par des tailles différentes selon le nombre de références qu'ils contiennent et la guilde est représentée par des couleurs différentes.

b. Guilde

Les espèces et genres ont souvent de fortes spécificités spatiales et géographiques, et afin de pouvoir comparer les communautés fongiques de différents sites ou environnements, les chercheurs ont recours à une classification en guildes trophiques. Les différents génomes ont été répartis en 4 guildes parmi : Saprotrophes, Ectomycorhiziens, Pathotrophes et Others, guildes pertinentes dans l'étude des sols, notamment forestiers.

Les pourcentages de bonnes prédictions à la guilde par taxons, par guilde, et en fonction de la plus proche référence taxonomique ont alors été calculés (figure 9). Ils sont meilleurs que les pourcentages de prédictions taxonomiques une médiane pour le boxplot de bonnes prédictions à la guilde de 86 % est observée pour les génomes dont le genre est présent dans la base de données. Cette médiane est plus basse lorsque la plus proche référence taxonomique des taxons s'éloigne du genre (on tombe à 40% lorsque la plus proche référence taxonomique est le phylum). La répartition des pourcentages de bons matchs à la guilde pour les différents taxons en fonction du nombre de référence est bien plus homogène que pour la taxonomie lorsque l'on considère des génomes dont la plus proche référence taxonomique n'est pas le genre. Enfin, la guilde Ectomycorhizienne semble mieux prédite dans l'ensemble que la guilde Saprotrophe ou Pathotrophe, lorsque le genre est présent dans la base de données. Cependant la prédiction pour cette guilde baisse considérablement lorsque la plus proche référence taxonomique est la famille ou l'ordre.

Il était intéressant par la suite de mesurer la fiabilité de prédiction de ces guildes pour les différents phylas de la base de données MycoCosm 2024. Ainsi les pourcentages de bonnes prédictions aux différentes guildes, pour les différents phylas, en fonction de nombre de références dans le taxon et de la plus proche référence taxonomique ont été calculés (figure 10). Les phylas *Ascomycota* et *Basidiomycota* obtiennent de bons pourcentages de bonnes prédictions à la guilde Saprotrophe (leurs médianes sont respectivement à 81% et 82%). Les prédictions pour la guilde Ectomycorhizienne sont moins bonnes pour le phylum *Ascomycota* (médiane à 64%) mais restent correctes pour le phylum *Basidiomycota* (médiane à 83%). Même observation pour les prédictions à la guilde *Pathotrophe* (médianes à 55% et 77% respectivement) mais le phylum *Ascomycota* obtient de bien meilleures prédictions à la guilde Others que *Basidiomycota* (médianes à 75% et 5% respectivement). Les taxons du phylum *Glomeromycota* obtiennent de bonnes prédictions pour la guilde "Others" (médiane à 84%) mais contiennent 2 références seulement. Le phylum *Mortierellomycota* qui montrait de mauvaises prédictions au genre obtient de très bonnes prédictions à la guilde saprotrophe (médiane à 97 %). Les génomes du phylum *Mucoromycota* sont très bien prédits pour la guilde Saprotrophe (médiane à 93%), et moins bien prédits pour la guilde

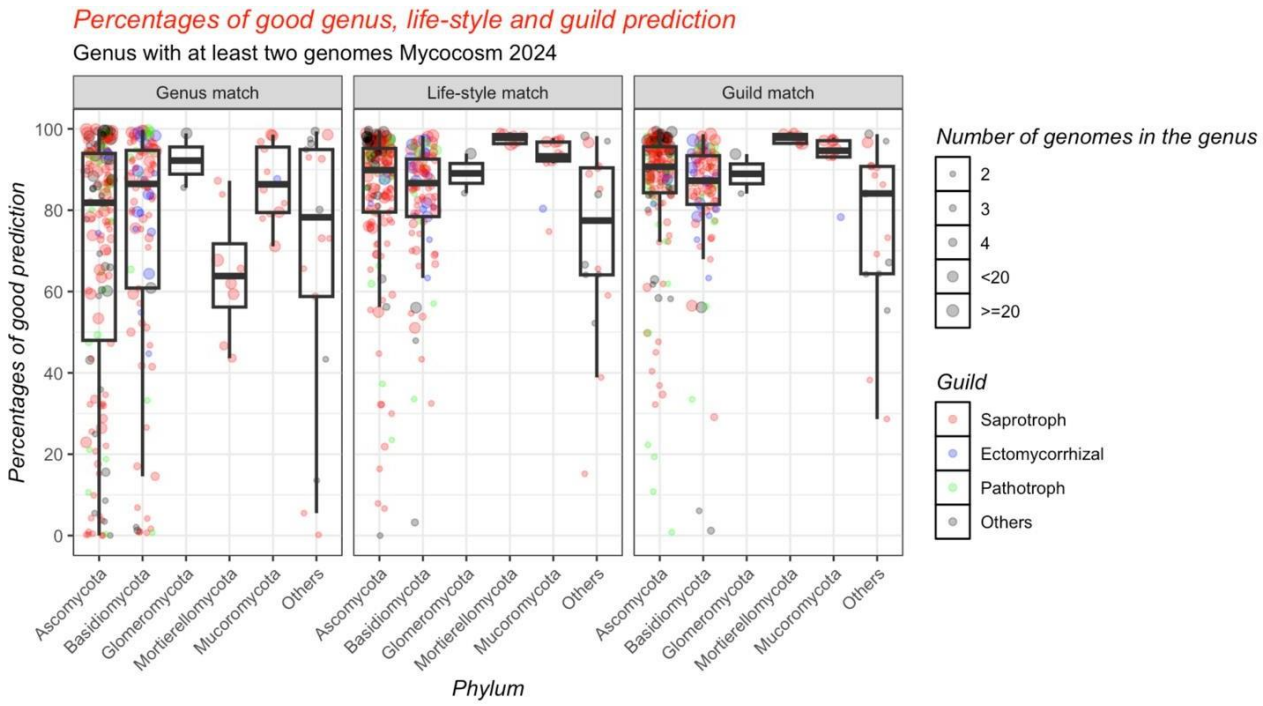

Figure 11 : Représentation boxplots des pourcentages des bons matchs, au style de vie et à la guilde pour les genres de tous les phylas possédant au moins une référence dans la base de données. Les genres sont représentés par des tailles différentes selon le nombre de références qu'ils contiennent et la guilde est représentée par des couleurs différentes.

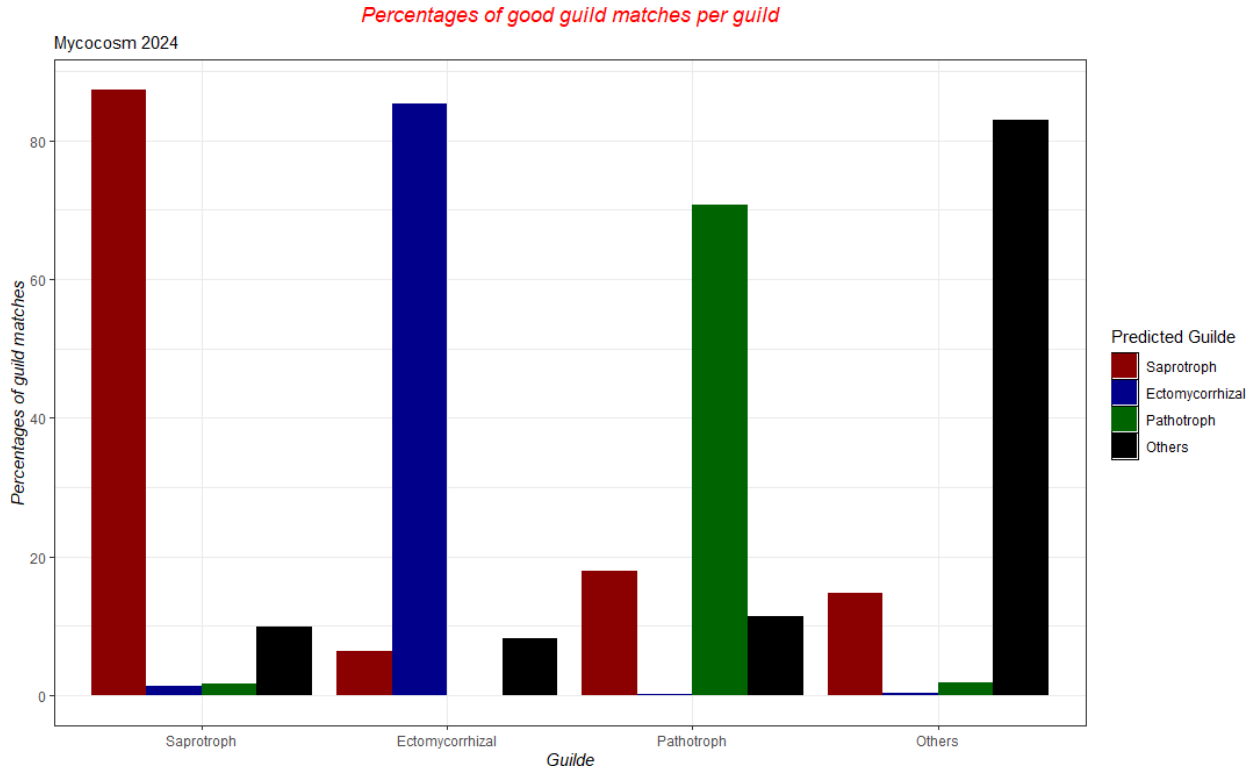

Figure 12 : Représentations histogramme des pourcentages de prédictions aux différentes guildes pour les guildes Saprotrophe, Ectomycorrhizienne, Pathotrophes et Others. Les différentes guildes sont représentées par différentes couleurs.

Ectomycorrhizienne (médiane à 78%). Enfin le phylum “Others” obtient des prédictions correctes pour la guilde Saprotrophe (médiane à 84%) et moins bonnes pour la guilde Others (médiane à 57%).

Les guildes ont été définies à partir des informations de style de vie primaire (32 modalités) et de style de vie secondaire (28 modalités) extraites à partir de la base de données Fungaltrait. La qualité de bonnes prédictions par genre à la guilde a ainsi été comparée à la qualité de bonnes prédictions au style de vie et à la qualité de bonnes prédictions au genre pour tous phylas (figure 11). Les prédictions à la guilde sont meilleures dans l’ensemble que les prédictions au style de vie qui sont meilleures que celles au genre.

Afin de mieux caractériser les conséquences de mauvaises prédictions, les proportions de prédiction des différentes guildes par les guildes Saprotrophes, Ectomycorhizienne, Pathotrophes et Others a été étudiée (figure 12). La guilde Saprotrophe est plutôt bien prédite comme Saprotrophe (87% de bonnes prédictions) et elle est prédite à tort comme guilde Others à 10%. La guilde Ectomycorhizienne est correctement prédite comme Ectomycorhizienne, à 85%. Elle est ensuite prédite à tort comme Others (8%) mais aussi comme Saprotrophe (6%). La guilde Pathotrophe obtient des pourcentages de bonnes prédictions à elle-même moins probants (70%), et match considérablement à tort sur la guilde Saprotrophe (18%) et Others (11%). La guilde Others obtient un pourcentage de matchs avec elle-même de 83% mais présente une certaine proportion de match avec la guilde Saprotrophe (15%).

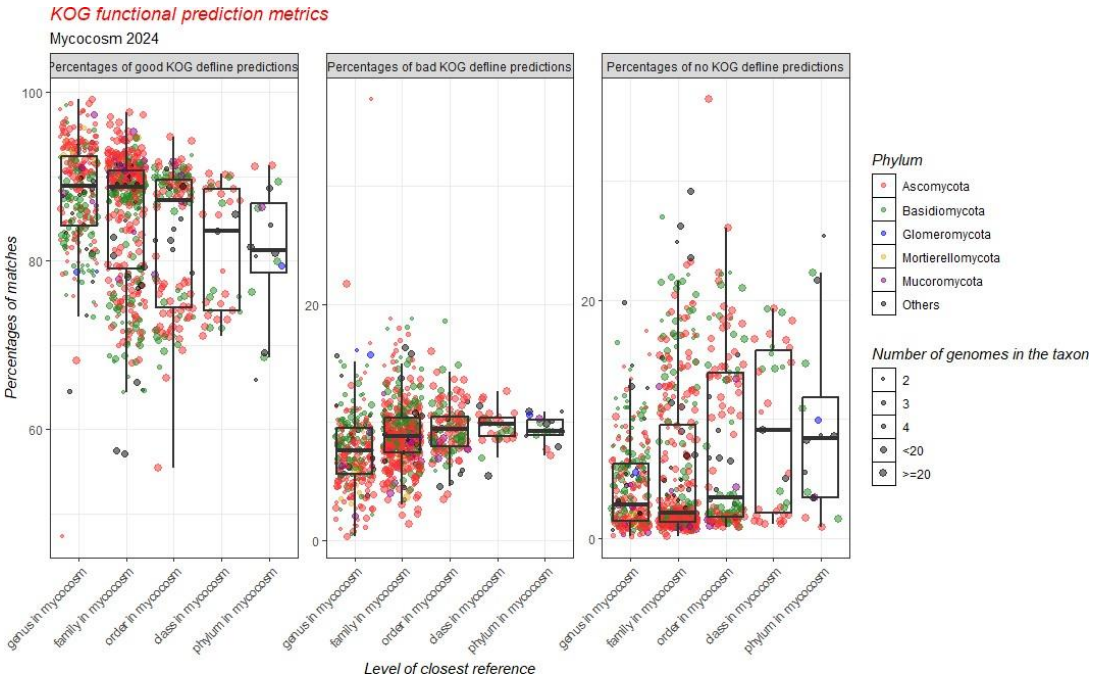

Figure 13 : Représentation boxplots des pourcentages de bonnes, mauvaises et non prédictions fonctionnelles au KOG defline par taxon, pour tous les phylas, en fonction de la plus proche référence taxonomique et du nombre de références par taxons. Les taxons sont représentés par des tailles différentes selon le nombre de références qu'ils contiennent et les différents phylas par des couleurs différentes.

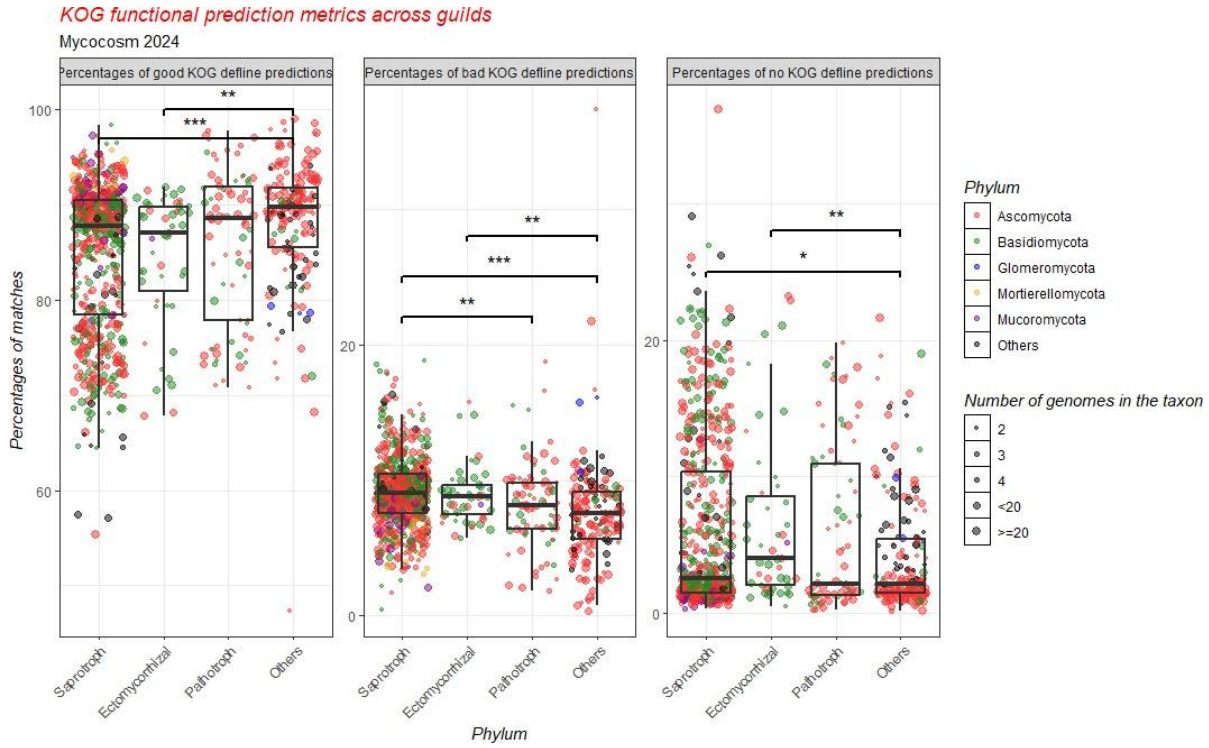

Figure 14 : Représentation boxplots des pourcentages de bonnes, mauvaises et non prédictions fonctionnelles au KOG defline par taxon, pour tous les phylas, en fonction du nombre de références par taxons et de la guilde. Les taxons sont représentés par des tailles différentes selon le nombre de références qu'ils contiennent et les différents phylas par des couleurs différentes.

c. Fonctions

Les pourcentages de bonnes prédictions à la KOG defline (le niveau le plus précis d'annotation KOG), de mauvaises prédictions à la KOG defline et de non-prédictions à la KOG defline ont été calculés par taxons, en fonction de la plus proche référence taxonomique et du phylum (figure 13). Le pourcentage de bonnes prédictions à la KOG defline pour les genres possédant une référence dans la base de données est bon (médiane à 88%). Ce pourcentage se maintient pour les familles possédant une référence dans la base de données puis décroit selon que le la plus proche référence des génomes s'éloigne du genre et de la famille, pour atteindre une médiane de 82% pour les génomes dont seul le phylum est présent dans la base de données. Pour les genres possédant une référence dans la base de données, on observe une amélioration de la qualité de bonnes prédictions à la KOG defline lorsque le nombre de référence augmente dans le genre. Cette tendance tend à s'effacer lorsque l'on considère des génomes dont la plus proche référence taxonomique n'est pas le genre. Les pourcentages de mauvaises prédictions à la KOG defline augmentent en fonction de la plus proche référence taxonomique à des médianes inférieurs10%. Les pourcentages de non-prédiction sont plus faibles pour les génomes dont le genre, la famille et l'ordre ont des références dans la base de données (<3%), puis grimpent pour une plus proche référence à la classe et au phylum sans pour autant dépasser la barre des 10%.

Les données de métatranscriptomiques des sols étant souvent analysées à l'échelle de la guilde trophique, les pourcentages de bonnes prédictions fonctionnelles ont été mesurés en fonction des différentes guildes (figure 14). Les médianes du pourcentage de bonnes prédictions à la KOG defline sont de 87, 86 et 88% pour les Saprotrophes, Ectomycorhiziens et Pathotrophe, sans différence significative. Seul le groupe "Others" présente des différences significatives avec les autres guildes. Le nombre de référence comprises dans les différents taxons ne semble pas influencer les pourcentages de prédictions à la KOG defline.

Les pourcentages de mauvaises prédictions à la KOG defline sont également plutôt constant pour toutes les guildes, entre 6% et 9% de médiane. Seules quelques comparaisons présentent des différences significatives, principalement avec le groupe "Others". Enfin les pourcentages de faux négatifs (pas de prédiction au KOG defline alors qu'une annotation est attendue), de 2 à 4%, ne présentent pas non plus de différences significatives, excepté pour le groupe "Others".

La prédiction de sécrétion, évaluée via la présence d'un Signal Peptides (SigP) qui induit la sécrétion des protéines qui le portent, est une annotation à très fort intérêt pour les chercheurs en écologie de sols, et l'évaluation de la qualité de cette prédiction a donc également été menée. Les pourcentages de vrais positifs (SigP attendu et prédit), faux positifs (SigP non attendu mais prédit),

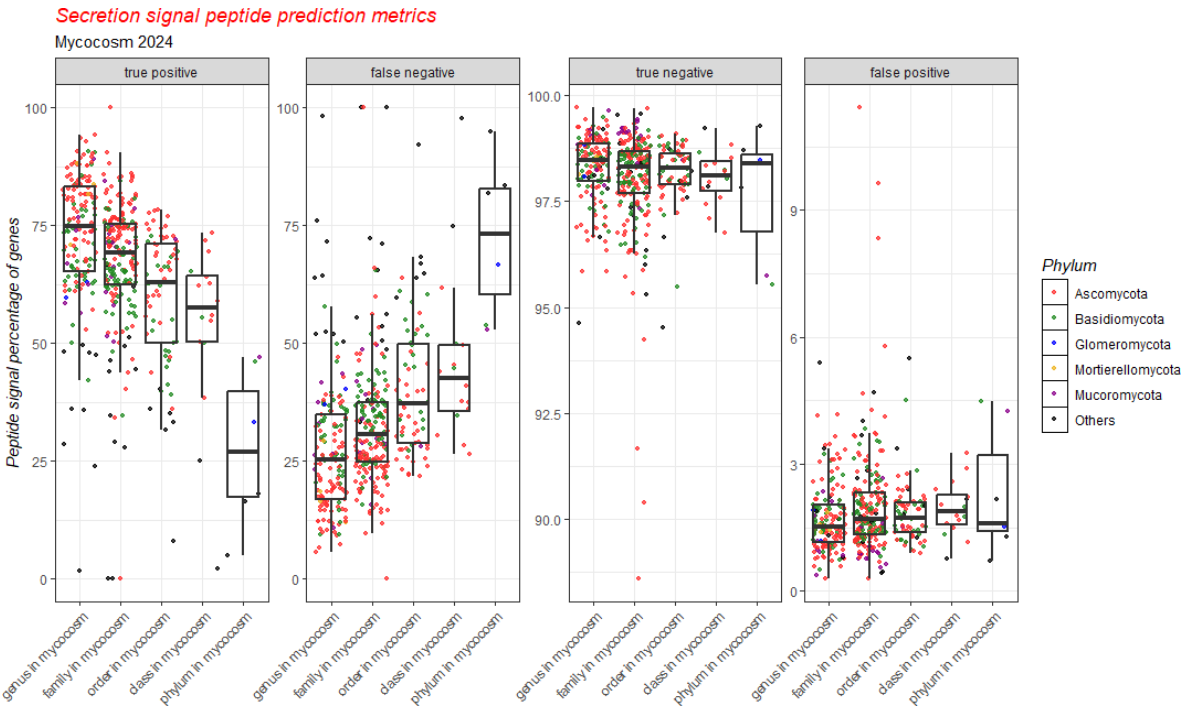

Figure 15 : Représentation boxplots des pourcentages de vrais positifs (SigP attendu et prédit), faux positifs (SigP non attendu mais prédit), vrais négatifs (SigP non attendu et non prédit) et faux négatifs (SigP attendu mais non prédit) pour la sécrétion de protéines pour tous les genres de la base de données, par phylas et en fonction de la plus proche référence taxonomique des genres. Les différents phylas sont représentés par des couleurs différentes.

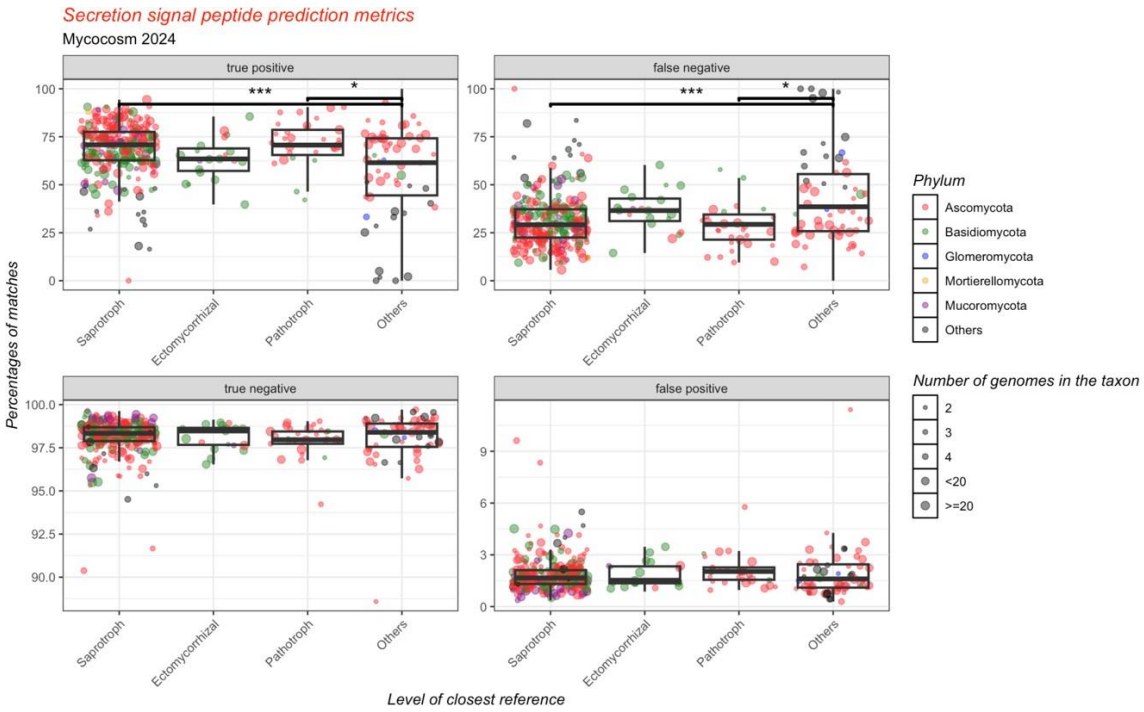

Figure 16 : Représentation boxplots des pourcentages de vrais positifs (SigP attendu et prédit), faux positifs (SigP non attendu mais prédit), vrais négatifs (SigP non attendu et non prédit) et faux négatifs (SigP attendu mais non prédit) pour la sécrétion de protéines pour tous les genres de la base de données, par phylas et en fonction des différentes guildes. Les différents phylas sont représentés par des couleurs différentes.

vrais négatifs (SigP non attendu et non prédit) et faux négatifs (SigP attendu mais non prédit) ont été mesurés pour tous les genres de tous les phylas de la base de données et en fonction de la plus proche référence taxonomique des genres (figure 15). Les pourcentages de vrais positifs sont plus faibles que les pourcentages précédemment obtenus sur les annotations KOG : on obtient ainsi une médiane à 75% de vrais positifs pour les genres possédant une référence au genre dans la base de données et ce pourcentage décroit rapidement lorsque l'on considère des genres dont la plus proche référence s'éloigne du genre (28% pour les genres ne possédant comme plus proche référence que le phylum). Les pourcentages de vrais négatifs sont meilleurs, à près de 99% lorsque les genres possèdent une référence au genre dans MycoCosm, et assez stables puisqu'ils se maintiennent au-dessus de 98% lorsque l'on considère des plus proches références taxonomiques autres que le genre. Les pourcentages de faux positifs sont pour la plupart des genres inférieurs à 5%, avec une médiane à 2%, assez stable quel que soit la proximité de la référence.

En considérant les guildes, malgré des variations de médiane, aucune différence significative n'est observée entre les pourcentages de vrais positifs des guilde Saprotrophe et Ectomycorhizienne et seul le groupe "Others" présente des différences significatives (figure 16). Il en va de même pour les pourcentages de faux négatifs, vrais négatifs et faux positifs, dont les distributions ne varient pas significativement entre les guildes d'intérêt.

4. Discussion :

MycoCosm est un portail de génomique fongique développé par le Joint Genome Institute (JGI) du ministère américain de l'énergie pour soutenir l'intégration, l'analyse et la diffusion de séquences de génomes fongiques et autres données "omiques", en particulier de celles du programme "1000 Fungal Genomes", qui vise notamment à combler les lacunes de l'arbre de vie fongique afin de résoudre d'importants problèmes liés à l'énergie et à l'environnement, en tirant parti des ressources croissantes de la génomique fongique (Grigoriev et al., 2014). L'organisation de cet arbre n'est pas du tout homogène, avec en particulier les phylas *Ascomycota* et *Basidiomycota*, regroupés dans les *Dikarya* qui est le groupe le plus important et diversifié du règne des Fungi. Ces deux phylas présentent de très nombreuses subdivisions en classes, ordres et familles, alors que les phylas extérieurs au Dikarya sont beaucoup moins diversifiés. La composition de Mycocosm suit naturellement cette structure, avec une très forte dominance des *Dikarya*. Par ailleurs, l'objectif du 1000 Fungal Genomes étant de couvrir au maximum la diversité des champignons, Mycocosm contient de nombreux génomes de champignons "exotiques" ou tout du moins atypiques, et de nombreux genres (plus de 600) n'y sont représentés que par un seul génome. Avec l'approche suivie dans cette étude de retirer des génomes pour les interroger contre le reste de la base de données, ces genres à 1 génome n'ont mécaniquement

pas pu conduire à une bonne annotation au genre. Mais dans un contexte d'échantillons réels, les génomes possédant une référence au genre dans notre analyse correspondent à des génomes appartenant à un genre qui contient deux références dans la base de données, et nos résultats sous-estiment donc probablement la qualité d'annotation pour des genres à faible nombre de génomes.

A l'inverse, certains genres tels *Aspergillus* ont de nombreuses références séquencées dans la base de données du fait de l'importance médicale ou économique qu'ils peuvent avoir, et des programmes de séquençages génomiques qui ont pu être lancé les concernant. Ces genres sur-représentés sont ainsi mieux prédits par l'approche basée sur les références dans MycoCosm et tendent à faussement "tirer vers le haut" les distributions des pourcentages de bonnes prédictions. Une analyse en moyenne par genre permet d'éviter ces biais et de donner le même poids sur la distribution à tous les genres, mais à l'inverse peut faire sous-estimer la qualité de prédiction en fonction des abondances et importances écologiques des taxons dans le cas de groupes moins hiérarchisés mais tout de même diversifiés. Par exemple, les *Glomeromycota* ont une faible diversité, mais deux genres avec un rôle écologique fondamental puisqu'ils forment des mycorhizes à arbuscules avec une énorme majorité des plantes terrestres. Dans Mycocosm, le phylum se divise en 2 ordres, 3 classes, 3 familles et 3 genres, mais il y a un très grand nombre d'espèces dans le genre Rhizophagus (10 génomes séquencés). Ce genre est très bien prédits, mais les autres beaucoup moins, en particulier *Geosiphon*, très atypique et différent dès le rang taxonomique de l'ordre. Les *Glomeromycota* n'apparaissent donc, en moyenne par genre, pas très bien prédits, alors qu'en contexte d'échantillon réel, la quasi-totalité des Glomerycota seraient des Rhizophagus et seraient très bien prédits.

L'analyse des cas de genres bien représentés mais présentant des pourcentages de bonnes prédictions faibles (au moins 3 références mais moins de 50% de bonne prédiction au genre) permet de proposer quelques pistes d'explications. Le genre *Motierella* du phylum *Mortierellomycota* a montré des pourcentages de bonnes prédictions au genre assez bas (65,74%), malgré 40 références présentes dans la base de données. Le genre *Mortierella* constitue un groupe ubiquiste très fréquemment isolé des sols et il est l'un de 10 genres fongiques les plus fréquemment retrouvés dans les projets de séquençage environnemental (Nagy et al., 2011). Les espèces de *Mortierella* sont largement répandues dans la zone tempérée, où elles sont presque cosmopolites en ce qui concerne plusieurs facteurs écologiques, se rencontrant dans une large gamme. Ce genre est un membre de la famille des *Mortierellaceae*, qui a longtemps été très mal résolue, c'est à dire dont la structure interne n'était pas claire. Elle a fait récemment l'objet d'un effort visant à la réduire en groupes monophylétiques (Vandepol et al., 2020), travail qui a conduit à la définition de 14 genres et à la réévaluation de la taxonomie de ses espèces. Ainsi le faible pourcentage de bonnes prédictions au genre pour *Mortierella* par la base de données MycoCosm peut être expliqué par plusieurs facteurs tels que sa

grande diversité, sa réévaluation taxonomique récente, sa complexité écologique et évolutive. Bien que 40 références soient présentes dans la base de données, cela peut ne pas être suffisant pour couvrir l'ensemble de la diversité génétique et écologique du genre *Mortierella*, et surtout de nombreuses espèces ont été par le passé faussement attribuées à ce genre. Les noms de genre-espèces n'ayant pas été modifiés dans Mycocosm et Fungaltraits, les faibles pourcentages de bonnes prédictions peuvent être expliqués par des erreurs dans la taxonomie des *Mortierellacae*. Les pourcentages de bonnes prédictions taxonomiques grimpent d'ailleurs rapidement pour le phylum dès que l'on observe des prédictions à la famille ou à un taxon plus large (plus de 95%).

Parmi les genres mal prédits, le genre *Erysiphe* qui présente un pourcentage de bonnes prédictions au genre de 47.56 % "match" 9444 fois avec lui-même, mais aussi 7256 fois avec le genre Oïdium. Il s'avère que ces deux genres sont en fait synonymes mais correspondent à des formes libres et formes pathogènes, historiquement séparées car elles n'avaient morphologiquement rien en commun et ont donc été identifiées comme des champignons différents. Même si les études phylogénétiques ont montré que les genres pouvaient être fusionnés, chez les pathologistes végétaux, les deux dénominations ont persisté. Pour d'autres genres, la synonymie n'est pas établie mais on retrouve des traits de vie très similaires voire identiques entre le genre attendu et le genre prédit. Par exemple, le genre *Thermothelomyces* "matche" sur lui-même à 32% (7686 gènes) mais "match" 14172 fois avec le genre *Myceliophthora*. Les espèces du genre *Thermothelomyces* sont thermophiles et présentent une croissance optimale à 45°C, tout comme les espèces du genre *Myceliophthora*.

Enfin, pour d'autres genres comme *Meliniomyces*, qui obtient un pourcentage de gènes correctement prédits au genre de 17.59% seulement, les autres genres prédits présentent des taxonomies peu fiables. A l'inverse des Erysiphe qui avaient des synonymes avec des formes morphologiques différentes, ceux-ci, tel *Chalara* par exemple, sont polyphylétiques, c'est à dire regroupent des taxons qui ne sont pas en lien phylogénétiquement, et ont historiquement été rassemblés en un seul genre sur la base de critères morphologiques comparables. Ainsi les cas de mauvaises prédictions de genre semblent souvent pouvoir être attribuées à des faiblesses dans la classification taxonomique des champignons utilisée avec Mycocosm, et ne remettent donc pas en cause la validité de l'approche. La qualité de prédiction taxonomique et ainsi meilleure qu'elle n'apparait, et les queues de distribution sont sans doute artéfactuelles.

Les faux négatifs (noHits) n'ont pas d'impact dans le contexte d'analyses métatranscriptomiques d'échantillons de terrain, car ils sont en général filtrés, et on souhaite donc surtout savoir si un gène pour lequel une annotation est prédite est correctement annoté. Un gène non prédit ne constitut pas une erreur au même titre qu'un gène mal prédit, et le taux d'erreur évalué en considérant le pourcentage de gènes avec la bonne annotation taxonomique est donc surestimé. Les pourcentages

de bonnes prédictions taxonomiques, trophiques et fonctionnelles ont été par la suite été calculés à partir de gènes ayant obtenus une prédiction uniquement, ce qui correspond mieux aux situations réelles. L'augmentation des pourcentages de bonnes prédictions taxonomiques qui résulte de ce changement de calcul est en miroir des pourcentages de noHits. Si on observe une augmentation 2 fois plus importante du pourcentage de bonnes prédictions à la taxonomie pour le phylum *Basidiomycota* que pour le phylum *Ascomycota,* c'est parce que le phylum *Ascomycota* présentait des pourcentages de faux négatifs 2 fois plus faibles que le phylum *Basidiomycota*. Les guildes trophiques sont dans l'ensemble mieux prédites que les genres, ce qui permet de corriger les biais de prédictions liés aux problèmes d'annotations taxonomiques. De plus, dans le cadre de l'étude menée au laboratoire INRAE Champenoux, une mesure de la qualité de prédiction à la guilde est plus pertinente qu'une mesure de la qualité de prédiction à la taxonomie : les genres fongiques étant spatialement, ou géographiquement très variables, et une comparaison inter-sites ou entre environnements différents est difficile sur la base des genres. La guilde est donc un niveau d'analyse plus pertinent que le genre, tout du moins en ce qui concerne l'écologie des sols. Cependant, dans l'arbre des champignons, les guildes trophiques ne sont pas structurées en sous-groupes monophylétiques mais existent dans différentes branches de l'arbre phylogénétique, traduisant des apparitions et transitions indépendantes de type trophiques ou d'hôtes (Spatafora et al., 2017). La répartition des différentes guildes pour les différents phylas n'est par ailleurs pas très homogène dans la base de données. Les Ectomycorrhiziens sont assez mal prédits dans le phylum *Ascomycota,* ce qui s'explique par une moins bonne homogénéité à la guilde dans ce phylum. En effet, les genres Ectomycorrhiziens sont souvent isolés dans des familles saprotrophes chez les *Ascomycota*, alors que les *Basidiomycota* présentent une répartition aux guildes Saprotrophe, Ectomycorrhizien et Pathotrophe plus homogène.

Les pourcentages de bonnes prédictions à la guilde pour les genres présent dans la base de données comparés aux pourcentages de bonnes prédictions au life-style et au genre confirment une meilleure prédiction de MycoCosm pour les guildes, les pourcentages de bonnes prédictions pour les plus grands phylas (*Ascomycota* et *Basidiomycota*) sont meilleurs et les queues de distributions correspondant aux genres mal prédits sont moins importantes.

Les guildes Saprotrophes, Ectomycorhiziennes et Others sont correctement prédites dans l'ensemble (plus de 80%) à l'exception de la guilde Pathotrophe qui n'obtient que 70% de bons matchs. Cette dernière est prédite à tort comme guilde Saprotrophe à un pourcentage de 18% mais aussi comme guilde Others à 15% et Ectomycorrhiziennes à 6%. Cela s'explique en partie par une sur-représentation de génomes Saprotrophes au sein de la base de données (1910 génomes sur les 2519 sélectionnés pour l'analyse).

En ce qui concerne les fonctions, les pourcentages de bonnes prédictions à la KOG defline sont meilleurs que pour la taxonomie et les guildes. Ils diminuent en fonction de la plus proche référence taxonomique considérée, mais la médiane par genre se maintient au-dessus de 80% même au phylum, et surtout, on n'observe pas de queue de distribution avec des génomes très mal prédits. Les pourcentages de prédictions erronées à la KOG defline sont en dessous de 10% de médiane et sont assez stables en fonction du niveau du plus proche référence taxonomique, et ce sont surtout les pourcentages de non-prédictions à la KOG defline qui augmentent en fonction du niveau du plus proche référence taxonomique. Ainsi lorsque qu'une espèce présente un "closest_ref" éloigné, l'outil de prédiction Diamond blast couplé à la base de données MycoCosm aurait plus tendance à ne pas donner de prédiction à la KOG defline plutôt qu'une mauvaise prédiction.

L'analyse de la qualité des prédictions à la KOG defline par guilde montre quelques différences significatives entre guildes, mais principalement entre le groupe « Others » et les autres. En particulier, on n'observe pas de différence significative des pourcentages de bonnes, fausses et absences de prédiction entre les Ectomycorrhiziens et les Saprotrophes. L'expression des gènes de ces deux guildes étant particulièrement pertinente pour les chercheurs en écologie microbienne des sols, il est donc intéressant de constater que la méthode n'induit pas de biais de qualité de prédiction entre ces deux guildes, malgré la sur-représentation des génomes saprotrophes par rapport au plus petit nombre de génomes Ectomycorhiziens.

Les résultats de prédiction de la sécrétion des protéines sont moins bons que ceux de l'annotation KOG, avec une médiane à 75% pour les vrais positifs, même pour les genres avec plusieurs génomes dans Mycocosm. De plus, la qualité de prédiction se dégrade très vite avec l'éloignement de plus proche niveau de référence taxonomique. Cependant, les pourcentages de faux positifs (prédiction d'un peptide signal alors qu'il n'est pas attendu) sont très faibles, en dessous de 3%, et stables pour les différents niveaux de "closest-ref", ce qui signifie que la sécrétion de protéines est très rarement prédite à tort par la base de données. Biologiquement, ces faibles pourcentages de bonnes prédictions des signal-peptides ne sont pas surprenants, car évolutivement le caractère sécrété d'une enzyme n'est pas extrêmement conservé, et les enzymes sécrétées peuvent être assez spécifiques d'une espèce. Par ailleurs, l'analyse des peptides signal réalisée pour le moment est potentiellement dégradée et ses résultats sous-estimés. En effet, l'annotation des peptides signal des génomes de Mycocosm a été réalisée avec l'outil SignalP, qui a évolué de 6 versions entre les cent premiers génomes du projet 1000 Fungal Genomes et les plus récents. Leur annotation n'est donc peut-être pas très homogène, et il pourrait être intéressant de réannoter les 2519 génomes de cette étude avec la même version de SignalP. Cette réannotation a été évaluée à une trentaine de jours de calculs sur 40 CPU, et n'a pas pu être réalisée dans le cadre de ce stage.

Les analyses fonctionnelles se sont concentrées sur les prédictions au KOG defline résumé par leurs pourcentages et par statut (bonne, erronée ou absente) dans un génome. Il s'agit de résultats encore préliminaires, et les chercheurs s'intéressent à la qualité de prédiction par type de fonction, par exemple, si les uréases, ou les transporteurs d'azote, sont correctement prédits, ou moins bien que la moyenne des gènes. Dans le laboratoire, ce type de comparaison entre des fonctions exprimées par les champignons Saprotrophes et Ectomycorrhiziens dans les écosystèmes forestiers au sein de différents sites a déjà été réalisée (Auer et al., 2023), mais demande validation. Les données générées dans ce projet contiennent l'information nécessaire pour répondre à ces questions, et cette analyse sera réalisée dans les prochaines semaines.

Les comparaisons Diamonds BlastX ont été réalisés à partir des fichiers CDS (coding sequence) de Mycocosm, donc transcrits complets de chaque gène. Ce n'est cependant pas la réalité d'échantillons deterrain séquencés : lors d'analyses métatranscriptomiques réelles sur des échantillons de sol, l'étape d'assemblage en contigs produit souvent des transcrits incomplets, et potentiellement des chimères résultantes de l'assemblage de morceaux de contigs provenant de différents ARNs. Afin d'évaluer l'effet de cette "incomplétude" sur les qualités de prédiction, une stratégie de dégradation *in silico* de la complétude des transcrits pourrait être envisagée. En testant plusieurs niveaux de complétude et en les comparant, l'effet des transcrits incomplets pourrait être mesuré, afin d'évaluer les qualités de prédiction des annotations taxonomiques et fonctionnelles en conditions non idéales.

Enfin, les Diamonds Blasts n'ont pas été réalisés avec une valeur de e-value standard, mais avec une valeur plus stringente, utilisée précédemment au laboratoire pour se prémunir de mauvaises annotations (10^{-10} au lieu de 10^{-3}). Cependant, il serait intéressant d'exploiter la même démarche que celle développée ici pour mesurer l'effet du paramétrage du Diamond Blast. L'hypothèse est qu'un seuil d'e-value plus élevé devrait faire diminuer le taux de "noHit", mais potentiellement au prix d'une augmentation des prédictions erronées, mais elle n'a jusqu'ici pas encore été testée dans le contexte de la prédiction d'annotation fongique avec Mycocosm.

5. Bibliographie :

H. Skovgård and T. Steenberg, 'Activity of Pupal Parasitoids of the Stable Fly Stomoxys Calcitrans and Prevalence of Entomopathogenic Fungi in the Stable Fly and the House Fly Musca Domestica in Denmark', *BioControl*, 47.1 (2002), pp. 45–60, doi:10.1023/A:1014434004946.

Erica Sterkenburg and others, 'Changes in Fungal Communities along a Boreal Forest Soil Fertility Gradient', *New Phytologist*, 207.4 (2015), pp. 1145–58, doi:10.1111/nph.13426.

Lynne Boddy, 'Chapter 11 - Fungi, Ecosystems, and Global Change', in *The Fungi (Third Edition)*, ed. by Sarah C. Watkinson, Lynne Boddy, and Nicholas P. Money (Academic Press, 2016), pp. 361–400, doi:10.1016/B978-0-12-382034-1.00011-6.

Leho Tedersoo, Mohammad Bahram, and Ian A. Dickie, 'Does Host Plant Richness Explain Diversity of Ectomycorrhizal Fungi? Re-Evaluation of Gao et al. (2013) Data Sets Reveals Sampling Effects', *Molecular Ecology*, 23.5 (2014), pp. 992–95, doi:10.1111/mec.12660.

Karolina Jörgensen and others, 'Ectomycorrhizal Fungi Are More Sensitive to High Soil Nitrogen Levels in Forests Exposed to Nitrogen Deposition', *New Phytologist*, 242.4 (2024), pp. 1725–38, doi:10.1111/nph.19509.

Lingtong Quan and others, 'Ectomycorrhizal Fungi, Two Species of *Laccaria*, Differentially Block the Migration and Accumulation of Cadmium and Copper in *Pinus Densiflora*', *Chemosphere*, 334 (2023), p. 138857, doi:10.1016/j.chemosphere.2023.138857.

Effects of *Rhizobium* Species Living with the Dark Septate Endophytic Fungus *Veronaeopsis Simplex* on Organic Substrate Utilization by the Host'
https://www.jstage.jst.go.jp/article/jsme2/33/1/33_ME17144/_article/-char/ja/

Abo Nouh and Fatma A, 'Endophytic Fungi for Sustainable Agriculture', *Microbial Biosystems*, 4.1 (2019), pp. 31–44, doi:10.21608/mb.2019.38886

Noemi Carla Baron and Everlon Cid Rigobelo, 'Endophytic Fungi: A Tool for Plant Growth Promotion and Sustainable Agriculture', *Mycology*, 13.1 (2022), pp. 39–55, doi:10.1080/21501203.2021.1945699.

Nhu H. Nguyen and others, 'FUNGuild: An Open Annotation Tool for Parsing Fungal Community Datasets by Ecological Guild', *Fungal Ecology*, 20 (2016), pp. 241–48, doi:10.1016/j.funeco.2015.06.006

C. P. Hawkins and J. A. MacMahon, 'Guilds: The Multiple Meanings of a Concept', *Annual Review of Entomology*, 34.Volume 34, 1989 (1989), pp. 423–51, doi:10.1146/annurev.en.34.010189.002231.

Minh-Phuong Nguyen and others, 'Host Species Shape the Community Structure of Culturable Endophytes in Fruits of Wild Berry Species (*Vaccinium Myrtillus* L., *Empetrum Nigrum* L. and *Vaccinium Vitis-Idaea* L.)', *FEMS Microbiology Ecology*, 97.8 (2021), p. fiab097, doi:10.1093/femsec/fiab097

Lucas Auer and others, 'Metatranscriptomics Sheds Light on the Links between the Functional Traits of Fungal Guilds and Ecological Processes in Forest Soil Ecosystems', *New Phytologist*, 242.4 (2024), pp.1676–90, doi:10.1111/nph.19471

R. Henrik Nilsson and others, 'Mycobiome Diversity: High-Throughput Sequencing and Identification of Fungi', *Nature Reviews Microbiology*, 17.2 (2019), pp. 95–109, doi:10.1038/s41579-018-0116-y.

Igor V. Grigoriev and others, 'MycoCosm Portal: Gearing up for 1000 Fungal Genomes', *Nucleic Acids Research*, 42.D1 (2014), pp. D699–704, doi:10.1093/nar/gkt1183

Morgan G. I. Langille and others, 'Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences', *Nature Biotechnology*, 31.9 (2013), pp. 814–21, doi:10.1038/nbt.2676.

Hazael Hernandez and Luis R. Martinez, 'Relationship of Environmental Disturbances and the Infectious Potential of Fungi', *Microbiology*, 164.3 (2018), pp. 233–41, doi:10.1099/mic.0.000620.

Natalie Vandepol and others, 'Resolving the Mortierellaceae Phylogeny through Synthesis of Multi-Gene Phylogenetics and Phylogenomics', *Fungal Diversity*, 104.1 (2020), pp. 267–89, doi:10.1007/s13225-020-00455-5.

Dylan P. Smith and Kabir G. Peay, 'Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing', *PLOS ONE*, 9.2 (2014), p. e90234, doi:10.1371/journal.pone.0090234.

Paul Bridge and Brian Spooner, 'Soil Fungi: Diversity and Detection', *Plant and Soil*, 232.1 (2001), pp. 147–54, doi:10.1023/A:1010346305799.

Salvador Lladó Fernández, Tomáš Větrovský, and Petr Baldrian, 'The Concept of Operational Taxonomic Units Revisited: Genomes of Bacteria That Are Regarded as Closely Related Are Often Highly Dissimilar', *Folia Microbiologica*, 64.1 (2019), pp. 19–23, doi:10.1007/s12223-018-0627-y.

Joseph W. Spatafora and others, 'The Fungal Tree of Life: From Molecular Systematics to Genome-Scale Phylogenies', in *The Fungal Kingdom* (John Wiley & Sons, Ltd, 2017), pp. 1–34, doi:10.1128/9781555819583.ch1.

Nicola T Case and others, 'The Future of Fungi: Threats and Opportunities', ed. by B Andrews, *G3 Genes|Genomes|Genetics*, 12.11 (2022), p. jkac224, doi:10.1093/g3journal/jkac224.

Amy Y. Rossman, 'The Impact of Invasive Fungi on Agricultural Ecosystems in the United States', in *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*, ed. by David W. Langor and Jon Sweeney (Springer Netherlands, 2009), pp. 97–107, doi:10.1007/978-1-4020-9680-8_7.

Richard B. Root, 'The Niche Exploitation Pattern of the Blue-Gray Gnatcatcher', *Ecological Monographs*, 37.4 (1967), pp. 317–50, doi:10.2307/1942327.

Martina Štursová and others, 'When the Forest Dies: The Response of Forest Soil Fungi to a Bark Beetle-Induced Tree Dieback', *The ISME Journal*, 8.9 (2014), pp. 1920–31, doi:10.1038/ismej.2014.37.

László G. Nagy and others, 'Where Is the Unseen Fungal Diversity Hidden? A Study of Mortierella Reveals a Large Contribution of Reference Collections to the Identification of Fungal Environmental Sequences', *New Phytologist*, 191.3 (2011), pp. 789–94, doi:10.1111/j.1469-8137.2011.03707.x.