Réalisation d'un pipeline d'analyse phylogénomique :

I. Introduction

L'étude des relations phylogénétiques chez les plantes est essentielle pour décrypter les mécanismes évolutifs ayant conduit à leur diversification et à leur adaptation à des environnements variés. Les plantes, l'un des groupes les plus diversifiés du règne vivant, ont vu leur évolution marquée par des événements majeurs, notamment les duplications génomiques entières (Whole Genome Duplication, WGD). Ces événements ont favorisé l'émergence de nouvelles fonctions génétiques et d'innovations évolutives, contribuant ainsi à la diversification des lignées végétales (Soltis et al., 2016). Cependant, ces dynamiques génomiques complexes, impliquant duplications, pertes et réarrangements, peuvent brouiller les signaux phylogénétiques, rendant difficile la distinction entre héritages partagés et événements indépendants, et complexifiant la reconstruction de l'histoire évolutive des espèces et de leurs gènes.

Pour surmonter ces défis, les approches de réconciliation phylogénétique constituent des outils puissants. En intégrant les phylogénies des gènes et des espèces, elles permettent d'identifier les duplications, les transferts horizontaux et les pertes tout en établissant des relations robustes entre les génomes.

Dans ce contexte, notre étude s'appuie sur une approche de réconciliation appliquée à des espèces cultivées représentatives des monocotylédones (blé, orge, riz, maïs) et des eudicotylédones (betterave, concombre, pastèque, carotte, soja, pommier, caféier, pomme de terre, poivron, chou), ainsi qu'à la plante modèle *Arabidopsis thaliana*. Ce travail vise à relier les données génomiques à l'histoire évolutive de ces espèces, en identifiant les signatures d'événements de duplication et en reconstruisant les trajectoires évolutives sous-jacentes.

II. Matériels et Méthodes :

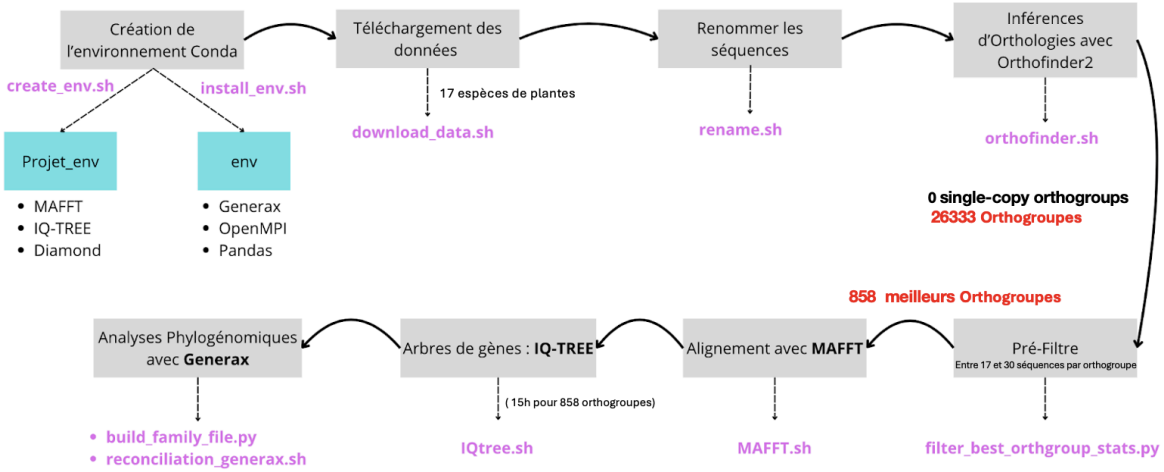


Figure 1 : Pipeline de l'analyse phylogénomique

1. Création de l'environnement

Dans le cadre de cette analyse, deux environnements ont été créés avec **Conda** : **Projet_env** et **env**, permettant de gérer efficacement les dépendances spécifiques à chaque étape du processus. L'environnement **Projet_env** inclut des outils essentiels pour l'alignement des séquences et les analyses phylogénétiques, tels que **Diamond**, **MAFFT**, et **IQ-TREE**. L'environnement **env** a été dédié à l'installation de packages utilisés dans d'autres analyses, tels que **Generax** pour l'inférence d'arbres phylogénétiques par méthodes bayésiennes, **OpenMPI** pour l'exécution parallèle, et **Pandas** pour la manipulation des données.

2. Téléchargement des données

Les données des espèces ont été extraites depuis la base de données EnsemblPlants (https://plants.ensembl.org/index.html). Le téléchargement des fichiers FASTA contenant les séquences protéiques des espèces sélectionnées a été automatisé en utilisant une commande **wget**. Le script ***download_data.sh*** a permis de télécharger les données, puis de les décompresser avec **gunzip** et de les organiser dans le dossier **raw_data.** Ce dernier *inclut un ensemble d'espèces végétales d'intérêt :* ***Amborella trichopoda*** *(outgroup)*, ***Arabidopsis thaliana*** (plante modèle), ***Beta vulgaris, Brassica oleracea, Capsicum annuum, Citrullus lanatus, Coffea canephora, Cucumis sativus, Daucus carota, Glycine max, Hordeum vulgare, Malus domestica, Marchantia polymorpha*** *(outgroup)*, ***Oryza sativa, Solanum tuberosum, Triticum aestivum, et Zea mays*** (Tableau I). L'inclusion des *outgroups* ***Amborella trichopoda*** et ***Marchantia polymorpha*** garantit une analyse complète et précise des événements évolutifs et d'établir des relations phylogénétiques robustes.

Tableau I : Liste des 17 espèces de plantes utilisés pour l'analyse phylogénomique

Nom scientifique	Nom commun	Caractéristique	Famille (Tribu)
Hordeum vulgare	Orge	Monocotylédone	Poaceae (Triticeae)
Triticum aestivum	Blé	Monocotylédone	Poaceae (Triticeae)
Oryza sativa	Riz	Monocotylédone	Poaceae
Zea mays	Maïs	Monocotylédone	Poaceae
Beta vulgaris	Betterave	Eudicotylédone	Amaranthaceae
Daucus carota	Carotte	Eudicotylédone	Apiaceae
Glycine max	Soja	Eudicotylédone	Fabaceae
Malus domestica	Pommier (Pomme)	Eudicotylédone	Rosaceae
Coffea canephora	Café	Eudicotylédone	Rubiaceae
Capsicum annuum	Piment / Poivron	Eudicotylédone	Solanaceae
Solanum tuberosum	Pomme de terre	Eudicotylédone	Solanaceae
Citrullus lanatus	Pastèque	Eudicotylédone	Cucurbitaceae
Cucumis sativus	Concombre	Eudicotylédone	Cucurbitaceae
Brassica oleracea	Chou	Eudicotylédone	Brassicaceae
Arabidopsis thaliana	Arabette des dames	Eudicotylédone	Brassicaceae
Amborella trichopoda	/	Angiosperme basal (Outgroup)	Amborellaceae
Marchantia polymorpha	/	Plantes non vasculaires (Outgroup)	Marchantiaceae

Afin de simplifier la suite de l'analyse, nous avons utilisé un script **rename.sh** afin de standardiser la nomenclature des espèces, en incluant le nom de l'espèce et son identifiant dans chaque en-tête de séquence, tout en supprimant les informations superflues dans les en-têtes afin de faciliter l'analyse pour les étapes suivantes. Il est essentiel de renommer les séquences afin d'assurer une identification claire et cohérente des données, ce qui permet de minimiser les erreurs pour la suite. Les fichiers renommés ont été stockés dans le répertoire **raw_data_rename**.

3. Inférences d'orthologies

L'identification d'orthologies a été établie avec l'outil **Orthofinder2** (Emms *et al.,* 2019). Grâce à cet outil, les gènes homologues issus de différentes espèces peuvent être regroupés en orthogroupes, ce qui facilite l'analyse des relations évolutives entre elles.
Le script orthofinder.sh a été utilisé pour automatiser le processus en exécutant OrthoFinder avec plusieurs threads afin d'accélérer le processus. Les résultats sont générés dans **orthoFinder_output** en traitant les fichiers d'entrée du répertoire **output_rename.** Les algorithmes **Dendroblast et DIAMOND** (Buchfink, *et al.,* 2015) sont utilisés par OrthoFinder pour effectuer un alignement rapide.
En sortie, OrthoFinder produit divers répertoires et fichiers, dont les plus importants pour cette étude sont les suivants :

- **Single_Copy_Orthologue_Sequences :** Liste des séquences de gènes orthologues à copie unique.
- **Orthogroup_Sequences :** Regroupe l'ensemble des séquences des orthogroupes repérés et partagés entre les diverses espèces (homologues et paralogues).

Le script Python **extract_summary_results_orthofinder.py** recueille et synthétise les statistiques globales d'OrthoFinder à partir du fichier **Statistics_Overall.tsv.** Il récupère des données essentielles comme le total de gènes, les gènes attribués à des orthogroupes, ainsi que d'autres métriques. Par la suite, les résultats sont enregistrés dans un fichier au format CSV (Tableau II).

4. Pré-Filtre des meilleurs orthogroupes (échantillonnage)

Le script Python **filter_best_orthgroup_stats.py** est utilisé pour filtrer les orthogroupes issus de l'analyse Orthofinder, ne retenant que les "meilleurs" orthogroupes. Concrètement, il applique un critère garantissant qu'il y a au moins une séquence par espèce dans chaque orthogroupe, soit autant de séquences que d'espèces (17), et un maximum de 30 séquences par orthogroupe. Ce seuil supérieur permet de prendre en compte les duplications génétiques ou génomiques, comme celles observées chez des espèces polyploïdes telles que le blé (hexaploïde). Ce filtrage assure ainsi que seules les orthogroupes les plus pertinentes et représentatifs de la diversité génétique des espèces étudiées sont retenues. Ce filtrage présente à la fois des avantages et des limites, qui seront discutés plus en détail dans la section des résultats.

Ce script est exécuté depuis un script bash launch_prefilter_orthologue.sh script copie les fichiers répondant aux critères dans EMP_filtered_orthogroups_min17_max30 et crée un fichier CSV orthogroups_stats_min17_max30.csv contenant des statistiques sur les orthogroupes filtrés, indiquant s'il a été retenu ou écarté pour chaque fichier, ainsi que le nombre de séquences et d'espèces qu'il contient.

5. Alignement avec MAFFT des séquences d'orthogroupes filtrées

Cette étape consiste à aligner les séquences des différents orthogroupes filtrés obtenus précédemment à l'aide de MAFFT, un outil d'alignement multiple de séquences (Kotah *et al.,* 2019). Le script MAFFT.sh prend en compte les fichiers FASTA (.fa) en effectuant un alignement multiple avec l'option **mafft-linsi**, et sauvegarde les résultats alignés dans le répertoire alignements_filtered_orthogroups_min17_max30 avec l'extension *_aligned.fa. L'alignement est effectué avec 4 threads pour accélérer le processus.

6. Arbres de gènes : IQtree

Après la réalisation des alignements, la prochaine étape est de créer des arbres phylogénétiques à partir des alignements de séquences en utilisant IQ-TREE (https://github.com/iqtree/iqtree2, & Nguyen *et al.,* 2015), un outil efficace pour inférer des arbres phylogénétiques. Chaque fichier d'alignement de l'analyse **MAFFT** est parcouru par le script IQtree.sh, qui utilise **IQ-TREE** pour établir un arbre phylogénétique en sélectionnant automatiquement le modèle de substitution le plus approprié en utilisant l'option **-m MFP**. L'option **-T** indique l'emploi de 4 threads afin de synchroniser le processus et de ralentir le calcul. La sortie du script contient les arbres phylogénétiques dans le répertoire « arbres ».

7. Analyses phylogénomiques

Plusieurs approches d'analyse phylogénomique, telles que les super-arbres, la concaténation et la réconciliation existent. L'approche de réconciliation a été retenue en raison de sa capacité à inférer des arbres d'espèces à partir de données génétiques complexes, en prenant en compte les événements de transfert horizontal de gènes, de duplication et de perte (DTL). Cette méthode permet de travailler directement de travailler à partir des orthogroupes en modélisant l'évolution des gènes (arbres de gènes). L'analyse a été réalisée à l'aide de **GeneRax** (https://github.com/BenoitMorel/GeneRax/blob/master/README.md , & Morel et al., 2020), et la réconciliation des arbres de gènes a été effectuée avec SpeciesRax (Morel et al., 2022), un outil intégré à GeneRax.

Le processus commence build_family_file.py, qui génère le fichier de famille contenant les informations nécessaires pour **GeneRax** à partir des alignements, arbres et d'autres données. Une fois le fichier de famille généré, le script reconciliation_generax.sh permet d'exécuter automatiquement GeneRax pour

inférer l'arbre des espèces à l'aide du modèle de réconciliation DTL. Ce script Bash inclut plusieurs paramètres importants pour réaliser cette réconciliation :

- **UndatedDTL** : Il s'agit du modèle de réconciliation utilisé pour modéliser les événements de duplication, de transfert et de perte de gènes. Le terme "Undated" signifie qu'aucune date n'est attribuée aux événements DTL, ce qui permet de ne pas tenir compte de la temporalité exacte de ces événements.
- **--strategy SKIP** : Ce paramètre indique que certaines étapes doivent être ignorées.
- **--si-strategy HYBRID :** Ce paramètre désigne une stratégie hybride dans laquelle plusieurs approches d'analyse sont combinées pour une meilleure estimation de la phylogénie.

Les résultats de ce script bash sont enregistrés dans **Output_generax**. Ce dernier contient plusieurs fichiers et sous-répertoires dans le répertoire resultats_speciesrax de cet output notamment **generax.log**, **fractionMissing.txt**, **stats.txt**, **perSpeciesCoverage.txt** etc.. Les répertoires **reconciliations, species_trees** et **results** sont les principaux points d'intérêt dans le répertoire de sortie.

- **reconciliations :** Contient les fichiers de la réconciliation des arbres de gènes en tenant compte des évènements de duplication, de perte et de transfert.
- **species_tree** : Ce répertoire contient les arbres des espèces générés par SpeciesRax, notamment **species_tree_quartet_support.newick**. Le **quartet support** est un type de support utilisé pour évaluer la solidité des branches. Il se base sur l'idée que l'arbre peut être divisé en sous-ensembles (quartets) de quatre espèces, et l'arbre est évalué sur la base du nombre de fois où chaque quartet spécifique est observé.
- **results** : Contient les résultats finaux de l'analyse dont les arbres réconciliés.

Les répertoires **reconciliations** avec les fichiers **".xml"** et **species_trees** avec les fichiers **".newick"** vont nous permettre d'obtenir les différents arbres de gènes et d'espèces, vont être utilisés avec des outils comme **SeaView** (Gouy *et al.,* 2021) et **RecPhyloXML** (http://phylariane.univ-lyon1.fr/recphyloxml/recphylovisu) pour visualiser et analyser les résultats. Les arbres obtenus seront présentés dans la partie résultats.

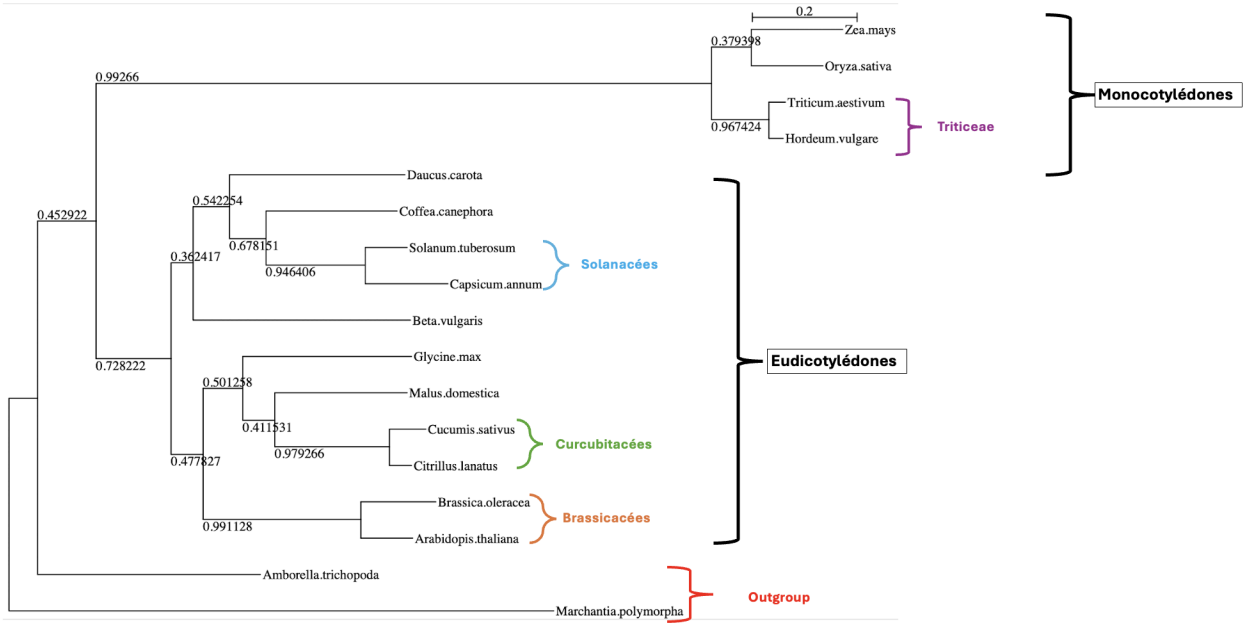
III. Résultats et Discussion :

Metric	Value
Total genes	797343
Genes assigned to orthogroups	747028
Percentage of genes assigned	93.69%
Total orthogroups	26333
G50	68
O50	3578
Orthogroups with all species	5114
Orthogroups with all species and single-copy genes	0

Tableau 2 : Résumé des Statistiques Globales de l'Analyse OrthoFinder

Statistiques globales de l'analyse OrthoFinder, incluant le nombre total de gènes, le nombre de gènes assignés à des orthogroupes, le pourcentage de gènes assignés, le nombre total d'orthogroupes, le nombre d'orthogroupes contenant au moins 50% des espèces (G50), le nombre d'orthogroupes contenant des gènes de plus de 50% des espèces (O50), le nombre d'orthogroupes contenant des gènes de toutes les espèces, et le nombre d'orthogroupes contenant des gènes à copie unique présents dans toutes les espèces.

Le tableau 2 résume les résultats obtenus avec Orthofinder. Parmi les 797 343 gènes analysés, 747 028 ont été attribués à des orthogroupes, soit 93,69 % du total. L'analyse a permis d'identifier **26 333 orthogroupes au total**. Le G50 est de 68, indiquant que 68 orthogroupes regroupent 50 % des gènes, tandis que l'O50 est de 3 578, signalant que 3 578 gènes couvrent 50 % des orthogroupes. De plus, 5 114 orthogroupes contiennent des gènes provenant de toutes les espèces étudiées. Toutefois, **aucun orthogroupe ne contient des gènes de toutes les espèces avec une copie unique par espèce**. Cela est dû à la fréquence des duplications génétiques chez les plantes, notamment la polyploïdie, qui entraîne la présence de copies supplémentaires de gènes. L'absence de gènes en copie unique par espèce empêche la réalisation d'un super-arbre ou d'une concaténation, et **la méthode choisie pour l'inférence a donc été la réconciliation**.



<u>Figure 2:</u> Arbre des espèces enracinés des plantes cultivés

Arbre phylogénétique enraciné issu de l'approche de réconciliation, illustrant les relations évolutives entre des espèces monocotylédones et dicotylédones. Les distances des branches reflètent les divergences évolutives, et ***Marchantia polymorpha*** *et* ***Amborella trichopoda*** *servent d'outgroups. L'échelle de l'arbre est de 0,2 substitutions par site, indiquant que pour chaque unité de longueur de branche, il y a 0,2 substitutions nucléotidiques en moyenne. Les valeurs sur les branches correspondent au* ***quartet support*** *qui permet d'évaluer la robustesse des branches.*

L'arbre phylogénétique enraciné illustre les relations évolutives entre des espèces monocotylédones et eudicotylédones, avec ***Marchantia polymorpha*** et ***Amborella trichopoda*** comme outgroups. Ces deux espèces, positionnées à l'extrémité de l'arbre, servent de référence pour orienter les relations évolutives des autres taxons. L'arbre met en évidence une séparation nette entre les deux grands groupes d'angiospermes : les monocotylédones (**blé, orge, riz, maïs**) et les eudicotylédones (**betterave, concombre, pastèque, carotte, soja, pommier, caféier, pomme de terre, poivron, chou**).Le **groupe des monocotylédones** est soutenu par un **quartet support élevé (0,99)**, tandis que celui des **eudicotylédones** présente un **support plus modéré** (**0,73**). Les branches des monocotylédones montrent des **longueurs élevées (environ 1,2 substitution par site)**, reflétant une divergence évolutive plus importante par rapport aux eudicotylédones.

Parmi les monocotylédones, ***Triticum aestivum*** (blé) et ***Hordeum vulgare*** (orge) forment un groupe frère avec une **faible distance évolutive post-divergence** et un **fort support de branche (0,96)**, en accord avec leur appartenance à la tribu des **Triticeae**. De manière similaire, ***Zea mays*** (maïs) et ***Oryza sativa*** (riz) apparaissent comme un groupe frère, mais avec un **support modéré (0,37)**, suggérant une relation évolutive moins certaine.

Chez les eudicotylédones, plusieurs relations proches se distinguent. Par exemple, ***Solanum tuberosum*** (pomme de terre) et ***Capsicum annuum*** (poivron), membres de la famille des **Solanacées**, forment un groupe frère avec un **support élevé (0,94)**, indiquant une divergence évolutive récente. Dans

la famille des **Cucurbitacées**, ***Cucumis sativus*** (concombre) et ***Citrullus lanatus*** (pastèque) forment également un groupe frère avec un **support élevé (0,97)**, illustrant leur proximité phylogénétique. De même, ***Arabidopsis thaliana*** et ***Brassica oleracea*** (chou), appartenant à la famille des **Brassicacées**, sont génétiquement très proches avec un **support quasi-parfait (0,99)**. En résumé, cet arbre montre clairement une **segmentation des monocotylédones et des eudicotylédones**, tout en révélant des relations évolutives précises au sein de chaque groupe. Les différences de support et de longueurs de branches reflètent les degrés de divergence et de robustesse des relations entre les espèces, mettant en évidence la complexité de l'évolution des angiospermes.

Pour optimiser les résultats obtenus avec GeneRax et renforcer la robustesse de l'arbre phylogénétique présenté dans la Figure 2, il aurait été préférable de **ne pas appliquer de filtre préalable sur les données**. Cela aurait permis d'exploiter pleinement la diversité et la richesse des séquences disponibles, en prenant en compte l'ensemble des événements évolutifs clés, à savoir les **duplications, transferts horizontaux et pertes de gènes**. Ces événements jouent un rôle essentiel dans l'évolution des gènes et leur réconciliation avec les arbres d'espèces. Cependant, ne pas appliquer de filtre sur les données aurait également conduit à un **temps de calcul considérablement augmenté**. À titre d'exemple, l'analyse de seulement 856 Orthogroupes a déjà nécessité environ 15 heures de calcul pour l'analyse IQTREE. Inclure l'ensemble des gènes disponibles sans filtrage aurait donc entraîné des temps de traitement bien plus longs, rendant difficile une inférence phylogénétique à grande échelle dans les délais disponibles.

Finalement, un exemple d'arbre de gènes a été généré à partir de l'orthogroupe OG0007488 (**Annexe 1**) en utilisant RecPhyloXML (http://phylariane.univ-lyon1.fr/recphyloxml/)**.** Les événements de spéciation sont bien définis, montrant des divergences claires entre les différentes espèces, comme la séparation entre les monocotylédones et les dicotylédones. Plusieurs événements de duplication sont observés, notamment dans les lignées de *Triticum aestivum* (blé) et *Oryza sativa* (riz), ce qui est cohérent avec la nature polyploïde de ces espèces.

IV. Conclusion :

Dans le cadre de cette étude, une approche de réconciliation a été appliquée aux espèces cultivées représentatives des monocotylédones et des eudicotylédones. L'objectif était de relier les données génomiques à l'histoire évolutive de ces espèces. En utilisant 858 des 26333 orthogroupes identifiés par OrthoFinder et la réconciliation des arbres de gènes avec GeneRax, un arbre phylogénétique détaillé a été construit, mettant en lumière les divergences évolutives entre les monocotylédones et les dicotylédones, et révélant des relations évolutives complexes au sein de chaque groupe.
Les résultats ont montré que les événements de duplication génétique sont omniprésents, notamment dans les lignées polyploïdes telles que *Triticum aestivum* (blé) et *Oryza sativa* (riz), ce qui a renforcé la pertinence de l'approche de réconciliation pour modéliser l'histoire évolutive de ces espèces. L'intégration des événements de duplication, transfert horizontal et perte de gènes a permis une meilleure compréhension des relations phylogénétiques entre les espèces étudiées.

Cependant, bien que l'application d'un filtre sur les données ait permis d'optimiser les temps de calcul, l'absence de ce filtre aurait permis une exploration plus large de la diversité génétique, renforçant ainsi la robustesse des résultats. En conséquence, cette étude a non seulement approfondi la compréhension des relations évolutives des espèces cultivées, mais a également mis en évidence les défis inhérents à l'analyse phylogénétique d'espèces présentant des duplications génétiques complexes. Cela souligne la nécessité d'adapter les méthodologies aux spécificités biologiques des espèces étudiées pour obtenir des résultats encore plus précis et pertinents.

V. Références :

Soltis, Pamela S., and Douglas E. Soltis. "Ancient WGD Events as Drivers of Key Innovations in Angiosperms." *Current Opinion in Plant Biology*, vol. 30, 2016, pp. 159–65, https://doi.org/10.1016/j.pbi.2016.03.015.

Emms, David M., and Steven Kelly. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology*, vol. 20, no. 1, Nov. 2019, p. 238, https://doi.org/10.1186/s13059-019-1832-y.

Buchfink, Benjamin, et al. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods*, vol. 12, no. 1, 2015, pp. 59–60, https://doi.org/10.1038/nmeth.3176.

https://www.ebi.ac.uk/jdispatcher/msa/mafft?stype=protein . Accessed 19 Dec. 2024

Katoh, Kazutaka, et al. "MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization." *Briefings in Bioinformatics*, vol. 20, no. 4, July 2019, pp. 1160–66, https://doi.org/10.1093/bib/bbx108.

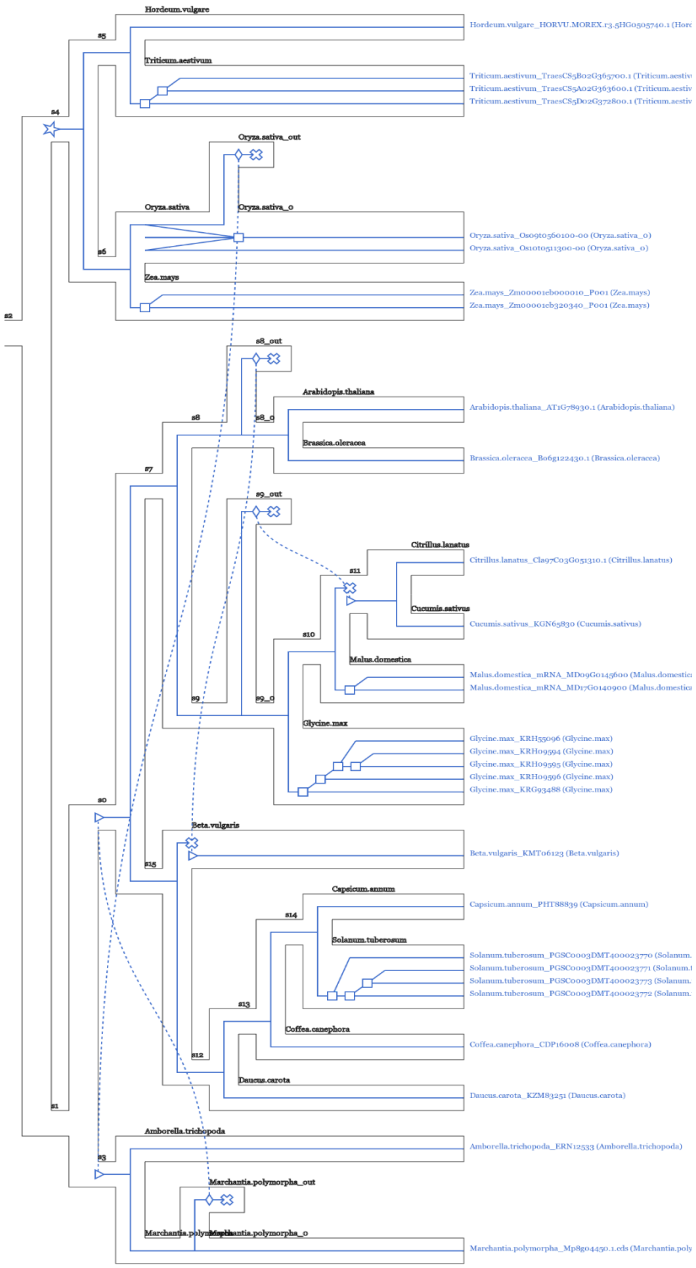
Nguyen, Lam-Tung, et al. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution*, vol. 32, no. 1, 2015, pp. 268–74, https://doi.org/10.1093/molbev/msu300.

Morel, Benoit, et al. "GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss." *Molecular Biology and Evolution*, edited by Rasmus Nielsen, vol. 37, no. 9, Sept. 2020, pp. 2763–74, https://doi.org/10.1093/molbev/msaa141.

Morel, Benoit, et al. "SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss." *Molecular Biology and Evolution*, edited by Tal Pupko, vol. 39, no. 2, Feb. 2022, p. msab365, https://doi.org/10.1093/molbev/msab365.

Gouy, Manolo, et al. "Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation." *Multiple Sequence Alignment*, edited by Kazutaka Katoh, vol. 2231, Springer US, 2021, pp. 241–60, https://doi.org/10.1007/978-1-0716-1036-7_15.

Annexes



Annexe 1 : Arbres de gènes* de l'orthogroupe OG0007488 *des espèces cultivées

Cette arbre illustre les relations évolutives entre différents gènes de diverses espèces végétales, mettant en évidence les événements de duplication et de spéciation.Les noeuds de l'arbre sont annotés avec des événements spécifiques tels que les duplications (D) et les spéciations (S), montrant des points de divergence évolutive