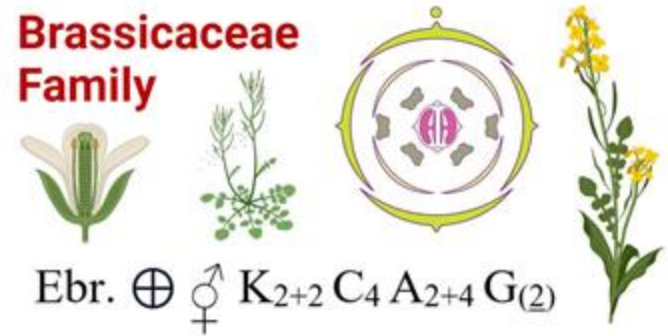
Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution

Chien-Hsun Huang,[1] Renran Sun,[1] Yi Hu,[2] Liping Zeng,[1] Ning Zhang,[3] Liming Cai,[1] Qiang Zhang,[4] Marcus A. Koch,[5] Ihsan Al-Shehbaz,[6] Patrick P. Edger,[7] J. Chris Pires,[8] Dun-Yan Tan,[9] Yang Zhong,[1] and Hong Ma*[1]



Molecular Biology and Evolution,
Volume 33, Issue 2, February 2016, Pages 394–412,
https://doi.org/10.1093/molbev/msv226
Published: 29 October 2015

Sirine Oueida
Valentin Goupille

M2 Bio-informatique
EMP

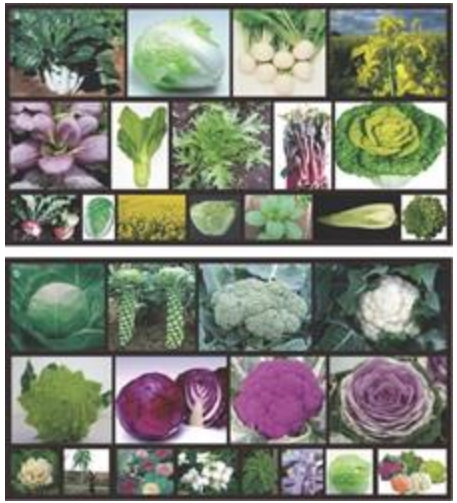
Université de Rennes

Introduction

Les Brassicacées

- L'une des familles d'angiospermes **les plus diversifiées** et **les plus rentable sur le plan économique**
  - **environ 3700 espèces** (51 tribus, 321 genres)

colza, betterave, choux, navet, brocoli, moutarde…



(Chen et *al,* 2014)

Modèle biologique végétale le plus étudié : ***Arabidopsis thaliana***

Introduction

Les Brassicacées

- **Une excellente famille modèle pour les études comparative et évolutives**
  - Diversité écologique, morphologique, génétique…
  - Histoire évolutive complexe → WGD (Whole Genome Duplication)

Pour faciliter l'utilisation des Brassicacées comme famille modèle évolutive
=> **une phylogénie avec des relations bien étayées est cruciale…**

2006 : Utilisation d'un **marqueur chloroplastique** chez 113 espèces de Brassicacées
=> ont identifié **3 lignées principales**
(avec des supports faibles à modérés)

Botany

Systematics and Phytogeography | Free Access

Brassicaceae phylogeny and trichome evolution†

Mark A. Beilstein, Ihsan A. Al-Shehbaz, Elizabeth A. Kellogg

First published: 01 April 2006 | https://doi.org/10.3732/ajb.93.4.607 | Citations: 308

Review

Cell

Cabbage family affairs: the evolutionary history of Brassicaceae

Andreas Franzke[1], Martin A. Lysak[2], Ihsan A. Al-Shehbaz[3], Marcus A. Koch[4] and Klaus Mummenhoff[5]

[1] Heidelberg Botanic Garden, Centre for Organismal Studies Heidelberg, Heidelberg University, D-69120 Heidelberg, Germany
[2] Department of Functional Genomics and Proteomics, Faculty of Science, Masaryk University, and CEITEC, CZ-625 00 Brno, Czech Republic
[3] Missouri Botanical Garden, St. Louis, MO 63166-0299, USA
[4] Biodiversity and Plant Systematics, Centre for Organismal Studies Heidelberg, Heidelberg University, D-69120 Heidelberg, Germany
[5] Biology Department, Botany, Osnabrück University, D-49069 Osnabrück, Germany

2011 : Le groupe de genres supplémentaires associés à LII avec un faible support a été appelé lignée étendue II (EII)

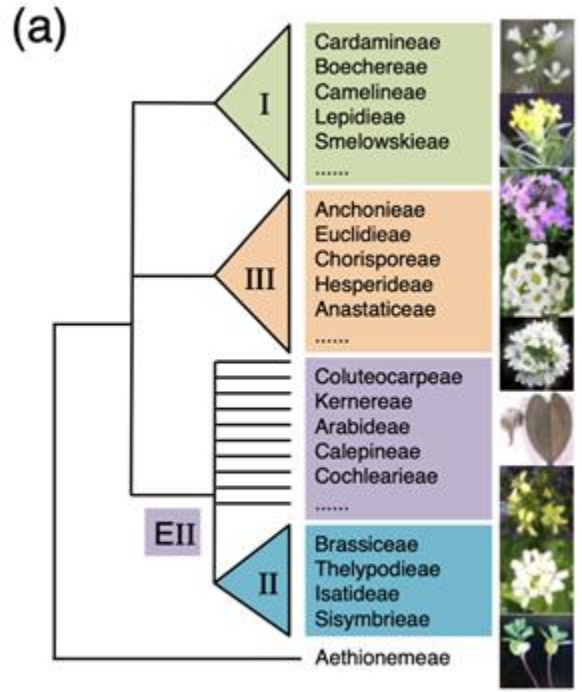


Figure : Phylogénie des Brassicacées
(Franzke et al, 2011)

Introduction

Les relations entre les principales lignées de la famille des Brassicacées restent floues…

Fournir une phylogénie des Brassicacées en utilisant des marqueurs nucléaires

(Huang et *al*, 2015)

Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution

Chien-Hsun Huang,[1] Renran Sun,[1] Yi Hu,[2] Liping Zeng,[1] Ning Zhang,[3] Liming Cai,[1] Qiang Zhang,[4] Marcus A. Koch,[5] Ihsan Al-Shehbaz,[6] Patrick P. Edger,[7] J. Chris Pires,[8] Dun-Yan Tan,[9] Yang Zhong,[1] and Hong Ma*[,1]

Objectifs :

I- Reconstruction phylogénétique des Brassicacées

II- Estimation des temps de divergences des Brassicacées

III- Reconstruction des caractères ancestraux à partir de traits morphologiques

I- Reconstruction phylogénétique des Brassicacées

Méthodologie :

1. Récupération et production des données génomiques / transcriptomiques
2. Sélection de gènes orthologues putatifs
3. Analyses Phylogénétiques
4. Évaluation de la robustesse phylogénétique et détection des biais

I- Reconstruction phylogénétique des Brassicacées

1) Récupération et production des données :

Cleome serrulata

- Production de 33 transcriptomes :
  - **32 espèces de Brassicaceae**
  - ***Cleome serrulata*, une Cleomaceae**
    - La famille sœur des Brassicaceae en tant que **groupe externe**

- Augmentées de **10 séquences de génomes entiers obtenues** à partir de bases de données publiques et de **13 ensembles de données de transcriptomes supplémentaires**

Un total de **55 grands ensembles de données** d'espèces de Brassicacées ont été utilisés **pour sélectionner un ensemble de gènes marqueurs pour les reconstructions phylogénétiques => couvrant 29 des 51 tribus de Brassicacées**

I- Reconstruction phylogénétique des Brassicacées

2) Sélection de gènes orthologues putatifs :

Reconstruction phylogénétique complexe chez les Brassicacées, compliquée par des événements de polyploïdisation, de pertes de gènes…

- Nécessité d'identifier des **gènes orthologues fiables** pour résoudre les relations entre lignées

=> **Trois approches ont été utilisées** pour sélectionner **113 gènes nucléaires à faible nombre de copies** comme marqueurs phylogénétiques…

I- Reconstruction phylogénétique des Brassicacées

3) Analyses phylogénétiques

A-Préparation des séquences

- **Obtention des gènes** : 113 gènes orthologues

- **Alignement des séquences** avec **MUSCLE**

- **Nettoyage des séquences** : élimination des séquences de mauvaises qualités et utilisation de **trimAl**

- **Concaténation** des séquences avec **SeaView**

  => Une **supermatrice** combinée de 113 gènes

I- Reconstruction phylogénétique des Brassicacées

B) Sélection du modèle de substitution :

- **GTRGAMMA :**

GTR (General Time Reversible) : Modèle qui permet des taux de substitution variables entre nucléotides

GAMMA : permet de prendre en compte les variations des taux de substitution entre sites

C) Utilisation de 2 méthodes de reconstruction phylogénétique :

- **Inférence Bayésienne** (BI) avec **MrBayes**
- **Maximum de Vraisemblance** (ML) avec **RAxML**

I- Reconstruction phylogénétique des Brassicacées

4) Évaluation de la robustesse phylogénétique et détection des biais

1. Évaluation de la confiance dans la sélection des arbres :

- **Test de 48 topologies** pour les relations entre les clades principaux de Brassicacées
- Analyse de robustesse avec **RAxML** et **CONSEL** (tests AU, bootstrap, PP, etc)

2. Détection des biais de séquence :

- Analyse des 113 gènes avec **TreSpEx** pour **détecter l'attraction des LB et la saturation**
- **Exclusion des gènes biaisés pour des reconstructions fiables**

3. Analyse de coalescence multi-espèces

- **Astral est** utilisé pour l'**analyse multilocus** basée sur les 113 arbres génétiques (Mirarab et al, 2014)

I- Reconstruction phylogénétique des Brassicacées

Concaténation

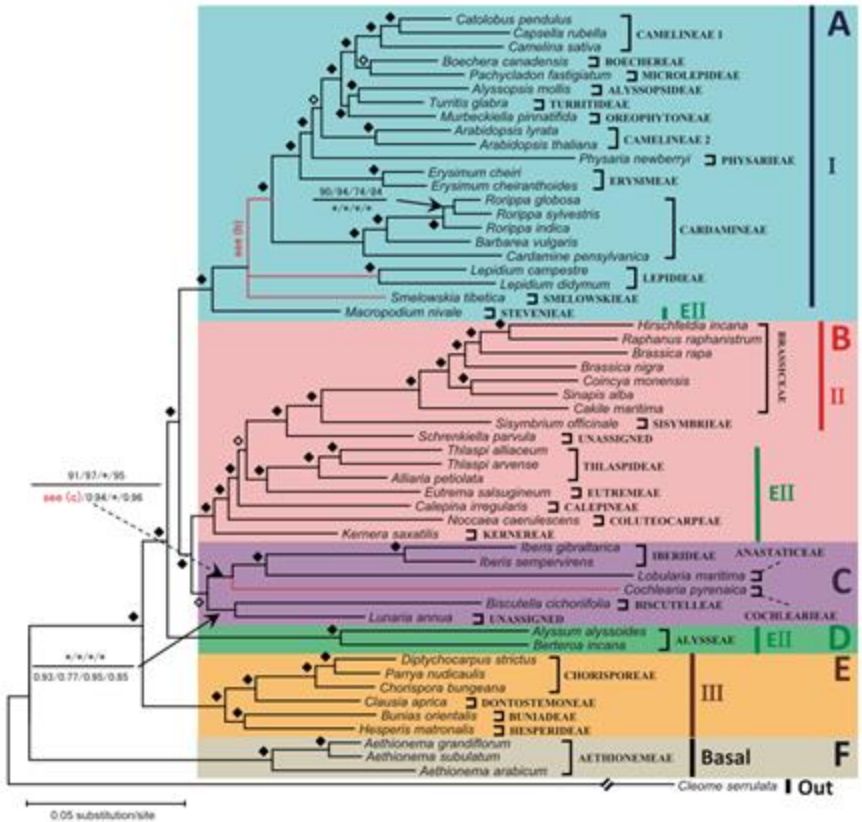
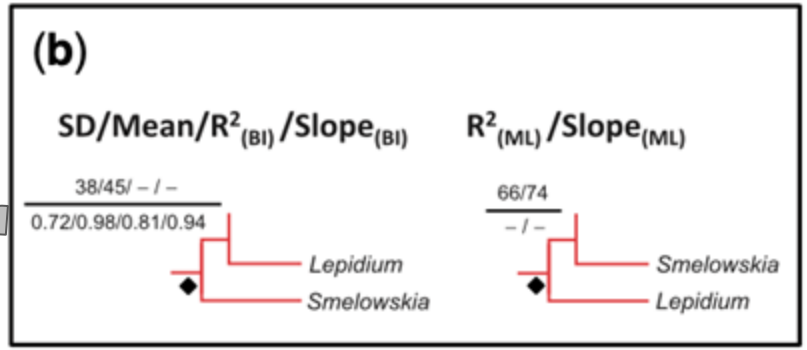


Figure 1 : Arbre phylogénétique des Brassicaceae basé sur les analyses ML et BI des concaténations d'ensembles de gènes après des tests de biais de séquence.

b) Relations entre *Smelowskia* et *Lepidium*



Smelowskia et *Lepidium* sont très proches phylogénétiquement dans l'arbre.

Différence entre les différentes méthodes statistiques utilisés

I- Reconstruction phylogénétique des Brassicacées

Consensus et Différence entre Concaténation et Coalescence

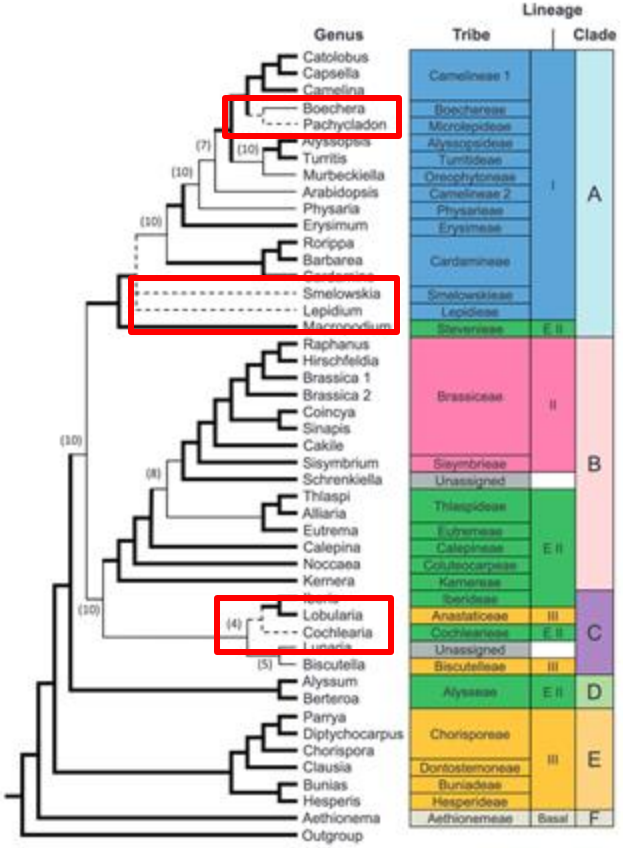


Figure 2 : Une phylogénie résumée des Brassicacées

Ce modèle est proposé selon nos résultats de **concaténation** (en utilisant 113 gènes comme marqueurs, et arbres après tests de biais et méthodes **coalescentes**

- Les **structures cohérentes en traits pleins**
  - Épais (toutes montrant un support maximal)
  - ou fins (avec des valeurs indiquant le nombre de résultats avec un support maximal).

- Les **lignes pointillées indiquent les incertitudes**

6 clades : A, B, C, D, E et F issues issus des méthodes de concaténation et de coalescence.

Les clades A, B et C → groupe (ABC) → groupe **monophylétique** → regroupé avec le clade D.

Clade E : Groupe séparé

Clade F *(Aethionemeae)* : lignée basale

I- Reconstruction phylogénétique des Brassicacées

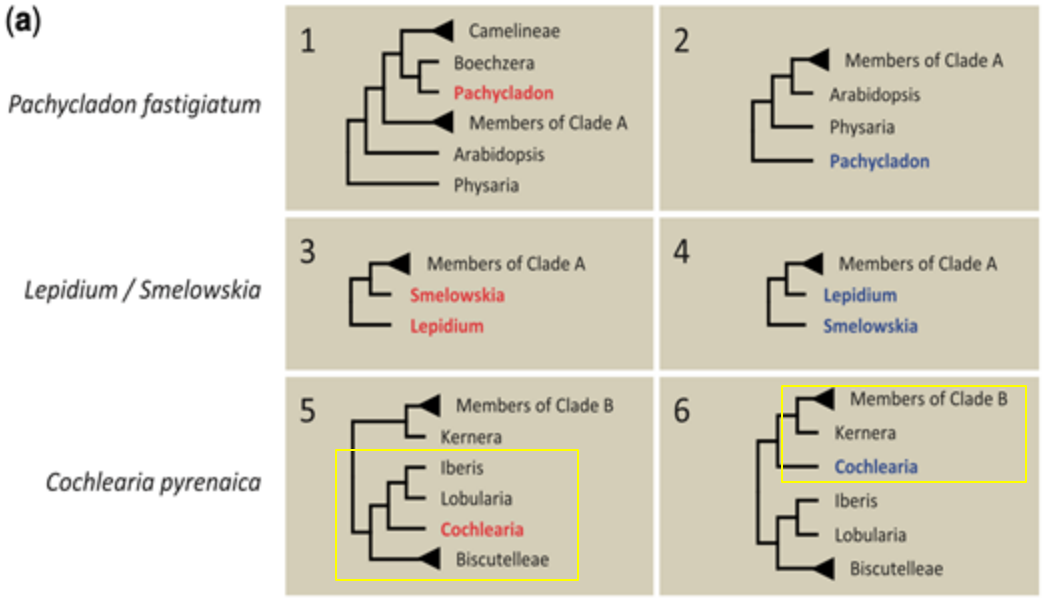


Figure 3 : Topologies conflictuelles de certains genres dans cette étude.

La position des genres peut varier en fonction des méthodes utilisées

Variations :
- Événements de réticulation
- Hybridation
- Polyploïdisation

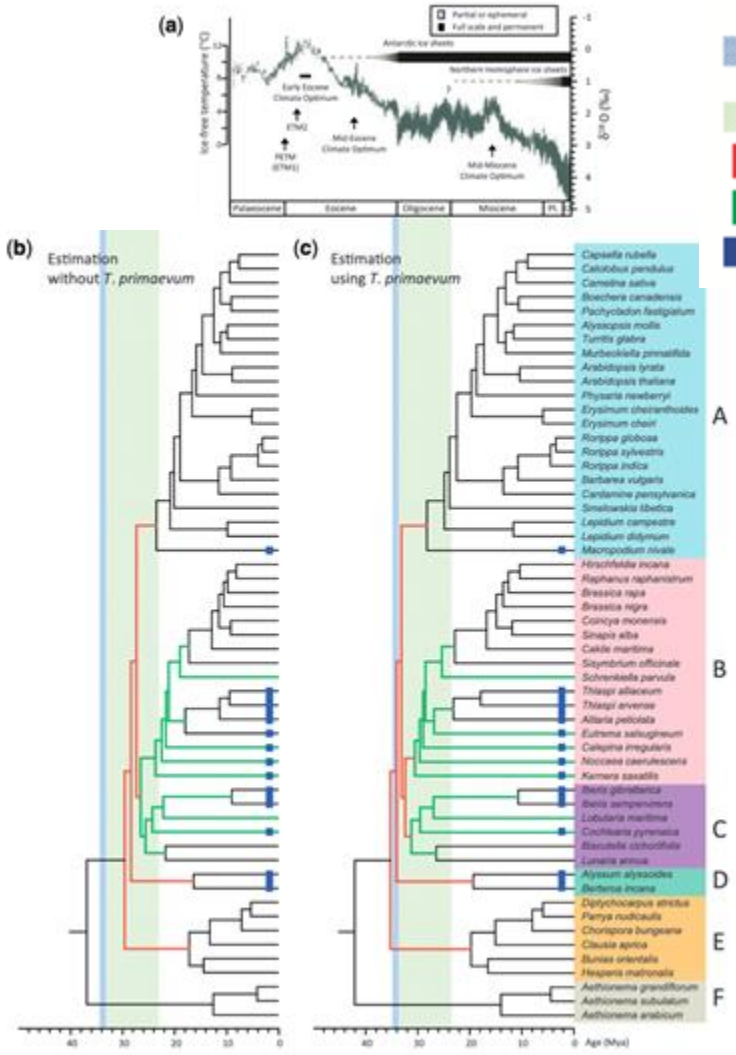
II- Estimation des temps de divergences des Brassicacées

Méthode utilisée : **Vraisemblance pénalisée (PL)** implémentée dans r8s

=> Hypothèse de taux variables d'évolution

L'arbre ML reconstruit à l'aide de 113 gènes avec une longueur de branche générée par RAxML a été utilisé comme arbre d'entrée

- **Calibration basée sur des fossiles**
  - ***D. bicarpelata* (89,3 Mya)** : Fossile utilisé pour calibrer l'âge du nœud de tige des Brassicales (incluant les Malvaceae comme groupe sœur).
  - ***T. primaevum* (23,03 Mya)** : Fossile supposé de Brassicaceae, utilisé pour contraindre l'âge minimum de la divergence entre *Thlaspi* et *Alliaria*.
  - **16 autres étalonnages** ont été utilisés en dehors du groupe des Brassicaceae pour estimer l'âge des groupes voisins
- **Estimations réalisées avec et sans fossile incertain (*T. primaevum*)**



Abrupt global cooling
(Eocene-Oligocene extinction event)

Oligocene

Early radiation of clades ABCDE

Nested divergences

EII tribes

II- Estimation des temps de divergences des Brassicacées

Figure 4 : Chronogramme des Brassicacées déduit à l'aide de r8s

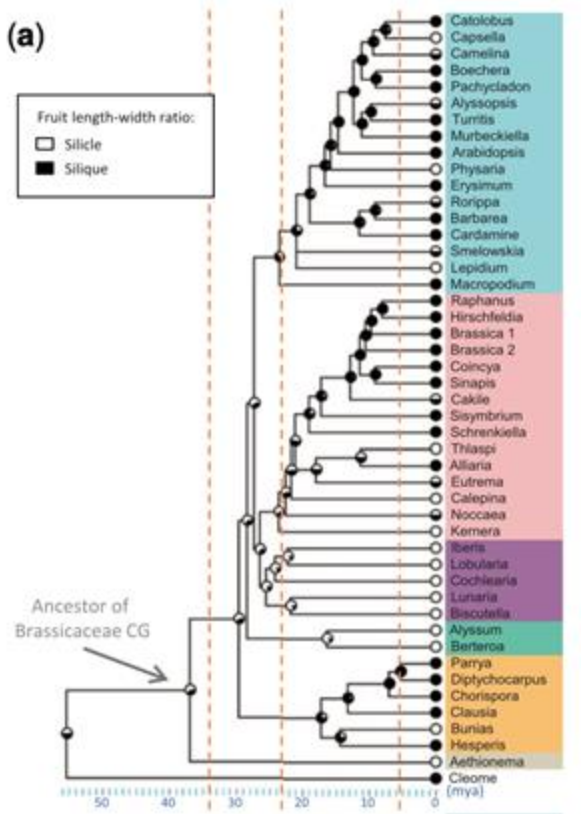
- **Concordance avec le paléoclimat** :
  - **Radiation des Brassicacées → période Eocène-Oligocène**

    correspondent aux changements climatiques du passé (refroidissement)

- **Impact du fossile** :
  - Inclusion ou non de ***T. primaevum*** : Résultats globalement cohérents : Différences minimes (~5 Ma) dans les estimations des temps de divergence

III- Reconstruction des caractères ancestraux à partir de traits morphologiques

Analyse des caractères morphologiques :



Reconstruction des états ancestraux pour plusieurs caractères morphologiques à l'aide de **Mesquite**

Silicule : Silique dont la longueur n'excède pas 3 fois la largeur

La reconstruction des caractères ancestraux révèle de nombreux événements d'évolution convergente

Figure 5 : Schémas évolutifs du rapport longueur-largeur des fruits

Conclusion

I- Reconstruction phylogénétique des Brassicacées

- Utilisation des marqueurs nucléaires
- Utilisation des méthodes de concaténation et de coalescence
- Une **phylogénie des Brassicacées largement confirmée**

→ Reste des incertitudes → **Incompréhension partielle** de la phylogénie globale

II- Estimation des temps de divergences des Brassicacées

- Changement climatique a joué un rôle dans la diversification des Brassicacées

III- Reconstruction des caractères ancestraux à partir de traits morphologiques

- La reconstruction des caractères ancestraux révèle de nombreux événements **d'évolution convergents**

Limites et perspectives

Limites :

- **Échantillonnage incomplet** : 19 tribus non représentées.
- **Incompréhension partielle** de la phylogénie globale.

Perspectives :

- **Espèces comme *Pachycladon* et *Cochlearia*** nécessitent plus de données pour clarifier leur **histoire évolutive**.
- **Augmenter l'échantillonnage** des espèces et des gènes pour **améliorer la compréhension** des relations évolutives.

Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity

Lachezar A. Nikolov[1], Philip Shushkov[2], Bruno Nevado[3], Xiangchao Gan[1], Ihsan A. Al-Shehbaz[4], Dmitry Filatov[3], C. Donovan Bailey[5] and Miltos Tsiantis[1]

[1]Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne 50829, Germany; [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA; [3]Department of Plant Sciences, University of Oxford, Oxford, OX1 3RB, UK; [4]Missouri Botanical Garden, 4344 Shaw Boulevard, St Louis, MO 63110, USA; [5]Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA

Author for correspondence:
Miltos Tsiantis
Tel: +49 22 15062105
Email: tsiantis@mpipz.mpg.de

Received: *9 August 2018*
Accepted: *10 January 2019*

New Phytologist (2019) **222:** 1638–1651
doi: 10.1111/nph.15732

Key words: anchored phylogenomics, comparative transcriptomics, crucifers, leaf evolution, targeted sequence capture.

Summary

• The Brassicaceae family comprises *c.* 4000 species including economically important crops and the model plant *Arabidopsis thaliana*. Despite their importance, the relationships among major lineages in the family remain unresolved, hampering comparative research.

• Here, we inferred a Brassicaceae phylogeny using newly generated targeted enrichment sequence data of 1827 exons (> 940 000 bases) representing 63 species, as well as sequenced genome data of 16 species, together representing 50 of the 52 currently recognized Brassicaceae tribes. A third of the samples were derived from herbarium material, facilitating broad taxonomic coverage of the family.

• Six major clades formed successive sister groups to the rest of Brassicaceae. We also recovered strong support for novel relationships among tribes, and resolved the position of 16 taxa previously not assigned to a tribe. The broad utility of these phylogenetic results is illustrated through a comparative investigation of genome-wide expression signatures that distinguish simple from complex leaves in Brassicaceae.

• Our study provides an easily extendable dataset for further advances in Brassicaceae systematics and a timely higher-level phylogenetic framework for a wide range of comparative studies of multiple traits in an intensively investigated group of plants.

(Nikolov et al., 2019, p. 1638)

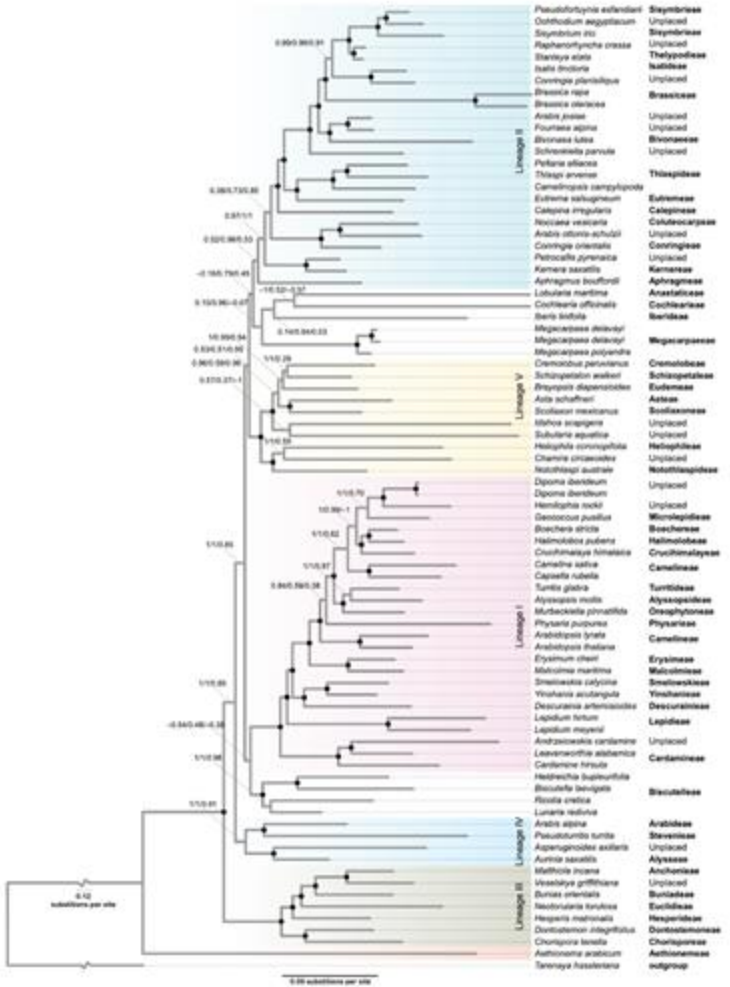


Fig. 3 Maximum-likelihood topology of the Brassicaceae relationships inferred from 79 species and 1421 exons. Nodes marked with a black circle are universally supported by all concatenation and coalescence analyses. Numbers associated with certain internodes are a quartet-based measure, lowest quartet internode certainty (LQ-IC) (Zhou et al., 2017), for quantifying the similarity between this reference tree and three sets of evaluation trees: RAxML bootstrap trees (2700 trees)/the RAxML bootstrap trees after removal of strong contributors (800 trees)/trees used to calculate local support in the coalescence analyses (2700 trees). LQ-IC varies between –1 and 1: it approaches ‘1’ when the reference internode is prevalent among the evaluation trees, it is close to ‘0’ when alternative topologies are common, suggesting incongruence between the reference tree and the evaluation trees with respect to this internode, and approaches ‘1’ when the evaluation trees strongly contradict the internode present in the reference topology. For example, the branch leading to Cremolobus peruvianus + Schizopetalon walkeri (1/1/0.28; Lineage V) is universally supported in both maximum-likelihood analyses (LQ-IC score = 1) but not present in all coalescent evaluation trees, reflected by a lower LQ-IC score (0.28).

CellPress
OPEN ACCESS

Current Biology

Article

Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset

Kasper P. Hendriks,[1,2,38,39,*] Christiane Kiefer,[3] Ihsan A. Al-Shehbaz,[4] C. Donovan Bailey,[5] Alex Hooft van Huysduynen,[2,6] Lachezar A. Nikolov,[7] Lars Nauheimer,[8] Alexandre R. Zuntini,[9] Dmitry A. German,[10] Andreas Franzke,[11] Marcus A. Koch,[3] Martin A. Lysak,[12] Óscar Toro-Núñez,[13] Barış Özüdoğru,[14] Vanessa R. Invernón,[15] Nora Walden,[3] Olivier Maurin,[9] Nikolai M. Hay,[16] Philip Shushkov,[17] Terezie Mandáková,[12] M. Eric Schranz,[18] Mats Thulin,[19] Michael D. Windham,[16] Ivana Rešetnik,[20] Stanislav Španiel,[21] Elfy Ly,[2,22,23] J. Chris Pires,[24] Alex Harkess,[25] Barbara Neuffer,[1] Robert Vogt,[26]

SUMMARY

The mustard family (Brassicaceae) is a scientifically and economically important family, containing the model plant *Arabidopsis thaliana* and numerous crop species that feed billions worldwide. Despite its relevance, most phylogenetic trees of the family are incompletely sampled and often contain poorly supported branches. Here, we present the most complete Brassicaceae genus-level family phylogenies to date (Brassicaceae Tree of Life or BrassiToL) based on nuclear (1,081 genes, 319 of the 349 genera; 57 of the 58 tribes) and plastome (60 genes, 265 genera; all tribes) data. We found cytonuclear discordance between the two, which is likely a result of rampant hybridization among closely and more distantly related lineages. To evaluate the impact of such hybridization on the nuclear phylogeny reconstruction, we performed five different gene sampling routines, which increasingly removed putatively paralog genes. Our cleaned subset of 297 genes revealed high support for the tribes, whereas support for the main lineages (supertribes) was moderate. Calibration based on the 20 most clock-like nuclear genes suggests a late Eocene to late Oligocene origin of the family. Finally, our results strongly support a recently published new family classification, dividing the family into two subfamilies (one with five supertribes), together representing 58 tribes. This includes five recently described or re-established tribes, including Arabidopsideae, a monogeneric tribe accommodating *Arabidopsis* without any close relatives. With a worldwide community of thousands of researchers working on Brassicaceae and its diverse members, our new genus-level family phylogeny will be an indispensable tool for studies on biodiversity and plant biology.

(Hendriks et al., 2023, p. 4052)

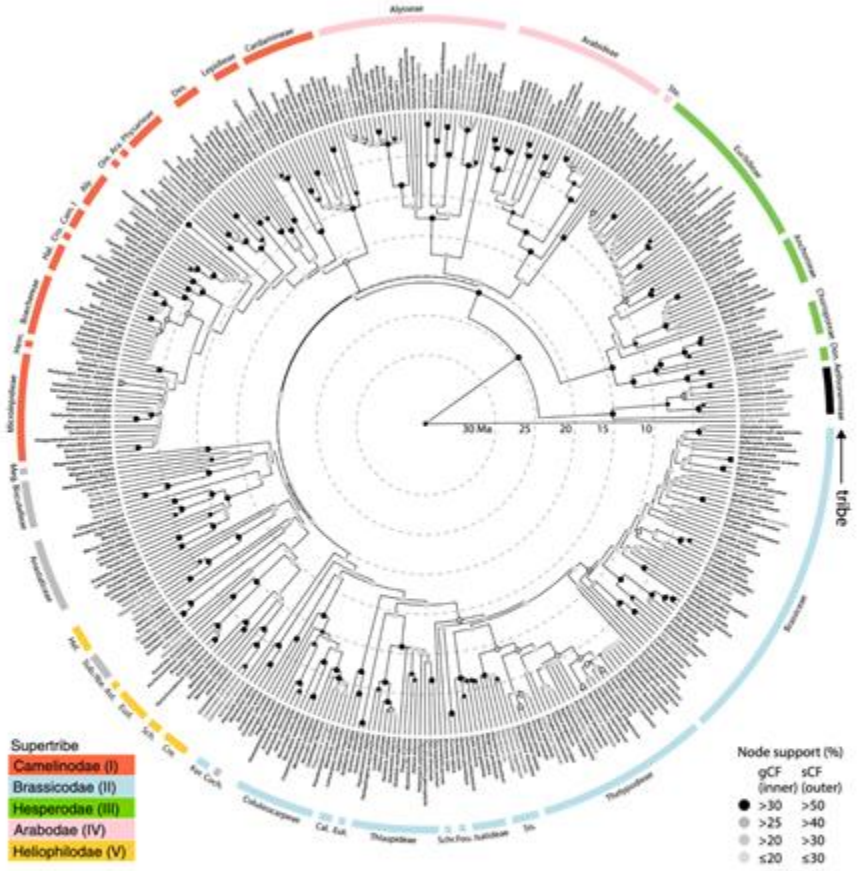


Figure 2. Time-calibrated genus-level Brassicaceae Tree of Life (BrassiToL) from a maximum likelihood analysis of a 297 nuclear genes supermatrix (superstrict routine)

Genus type species are highlighted in bold. All tribes with more than a single representative are listed. Abbreviations of tribes are as follows: Aly., Alyssopsideae; Ara., Arabidopsideae trib. nov.; Ast., Asteae; Cal., Calepineae; Cam. I, Camelineae I; Coch., Cochlearieae; Cre., Cremolobeae; Cru., Crucihimalayeae; Des., Descurainieae; Don., Dontostemoneae; Eud., Eudemeae; Eut., Eutremeae; Fou., Fourraeeae; Hal., Halimolobeae; Hel., Heliophileae; Hem., Hemilophieae; Ibe., Iberideae; Ker., Kernereae; Meg., Megacarpaeeae; Ore., Oreophytoneae; Sch., Schizopetaleae; Schr., Schrenkielleae trib. nov.; Sis., Sisymbrieae; Ste., Stevenieae; Sub., Subularieae. See also Data S2A for a fully annotated version of this phylogeny (including bootstrap, gCF, sCF and node age 95% HPD intervals) with outgroups representing all families within the order Brassicales and calibration nodes. See also Data S2A.

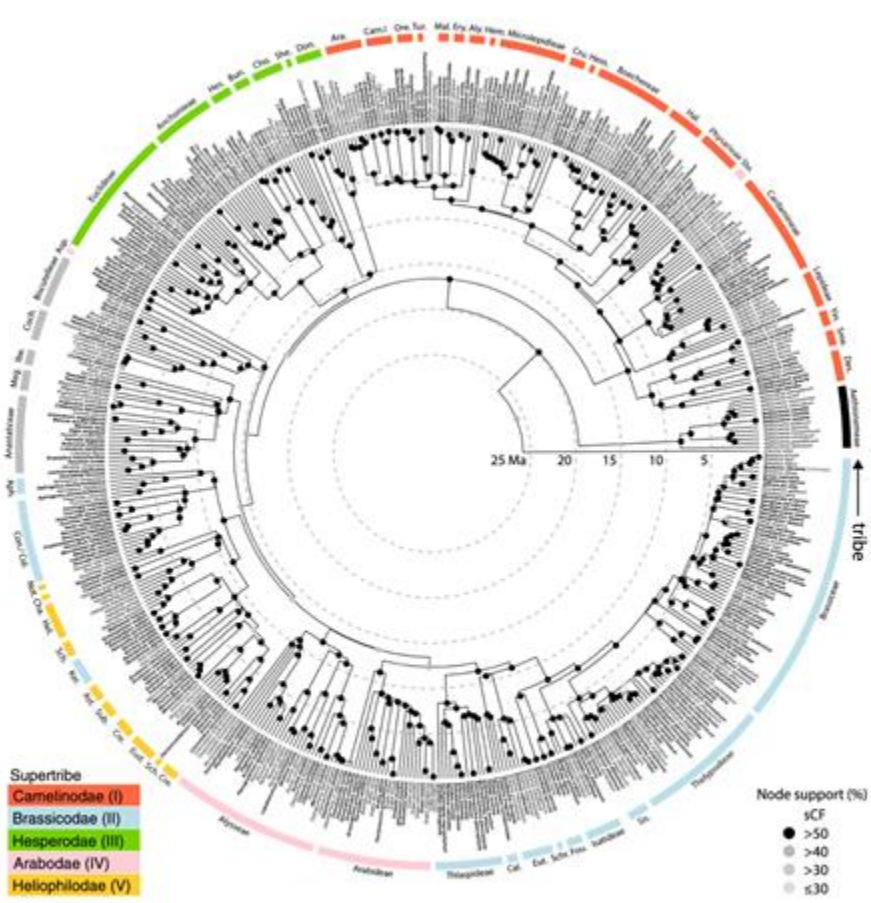


Figure 3. Time-calibrated genus-level Brassicaceae Tree of Life from a maximum likelihood analysis of a 60 plastome genes supermatrix

Genus type species are highlighted in bold. All tribes with more than a single representative are listed. Abbreviations of tribes follow those of Figure 2, with additionally: Aph., Aphragmeae; Asp., Asperuginoideae trib. nov.; Bun., Buniadeae; Cha., Chamireae; Cho., Chorisporeae; Col., Coluteocarpeae; Con., Conringieae; Don., Dontostemoneae; Ery., Erysimeae; Hes., Hesperideae; Mal., Malcolmieae; Not., Notothlaspideae; Ore., Oreophytoneae; She., Shehbazieae; Sme., Smelowskieae; Tur., Turritideae; Yin., Yinshanieae. See also Data S2B for a fully annotated version of this phylogeny (including bootstrap, sCF and node age 95% HPD intervals) with outgroups representing all families within the order Brassicales and calibration nodes. See also Data S1.

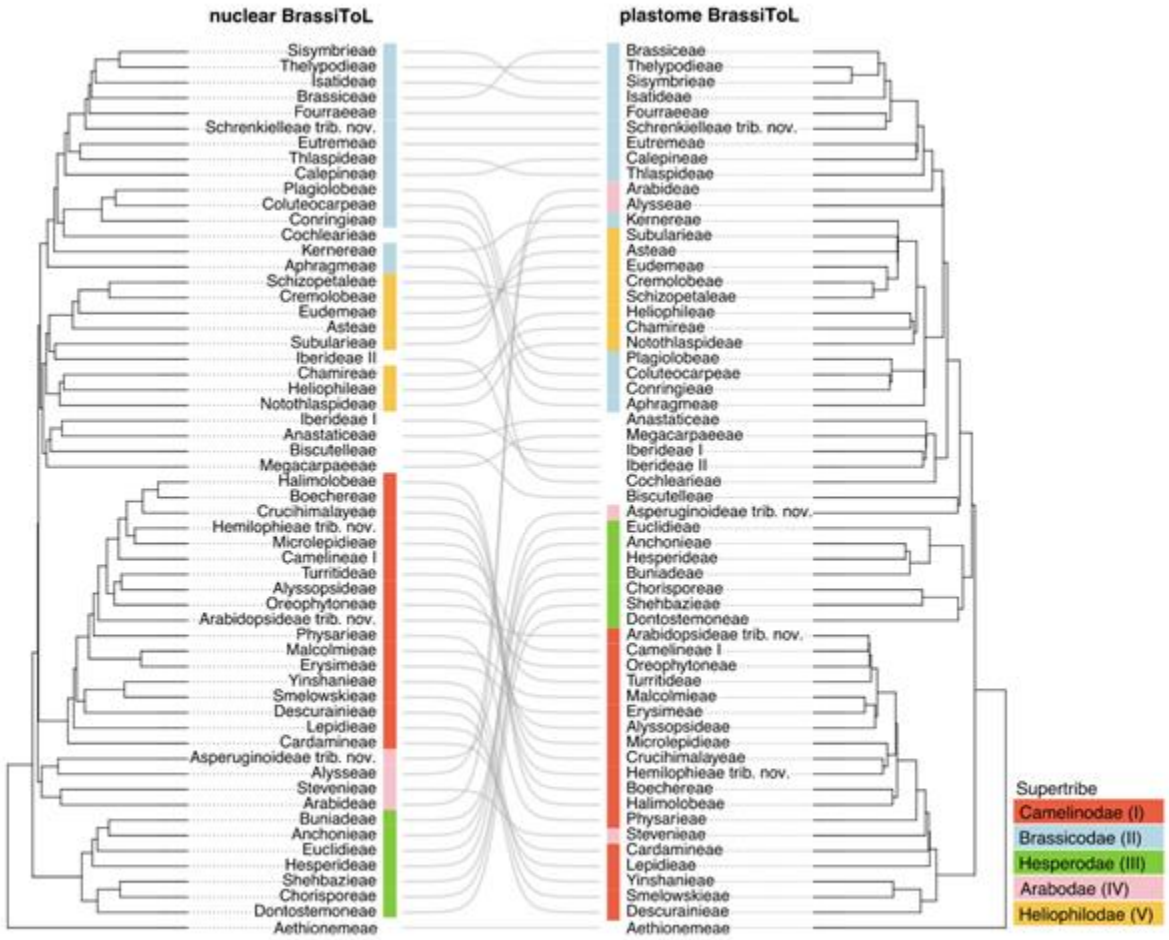


Figure 4. Cytonuclear discordance at tribe level in newly derived nuclear and plastome Brassicaceae Trees of Life

Curved lines between tip labels from the two phylogenies link the same tribes. Tribes are represented by a randomly chosen sample from each tribe. Rogue tribes have not yet been assigned to a supertribe due to their changing topological position from different sampling routines and phylogenetic approaches.

Table 2. Brassicaceae divergence time estimates

Studies	Crown Brassicaceae (Ma)	Crown core Brassicaceae[a] (Ma)	Method	Dataset	Calibration
Koch et al.[79]	–	25.9–23.1	synonymous substitution rate	*Adh* and *Chs*	synonymous substitution rate
Franzke et al.[18]	35.0–15.0–1.0	28.0–11.0–1.0	BEAST	*nad4*	one secondary calibration
Beilstein et al.[19]	64.2–54.3–45.2	54.3–46.9–39.4	BEAST	*ndhF* and *PHYA*	four fossils
Couvreur et al.[20]	49.4–37.6–24.2	43.8–32.3–20.9	BEAST	eight genes from nuclei, chloroplast, and mitochondria	one fossil
Kagale et al.[80]	–	26.6	synonymous substitution rate	213 nuclear orthologs	synonymous substitution rate
Edger et al.[11]	45.9–31.8–16.8	–	BEAST	1,115 single-copy nuclear genes	two fossils
Hohmann et al.[9]	38.6–32.4–27.1	27.3–23.4–19.9	BEAST	plastomes	four fossils
Huang et al.[22]	37.8–37.1–36.3	30.3–29.7–29.1	r8s	113 low-copy nuclear orthologs	18 fossils[b]
Cardinal-McTeague et al.[54]	44.1–37.7–31.4	–	BEAST	Chloroplast DNA (ndhF, matK, rbcL) and mitochondrial DNA (matR, rps3)	three fossils[b]
Mohammadin et al.[81]	58.9–48.0–37.5	35.4	BEAST	plastomes	one secondary calibration
Guo et al.[82]	41.8–34.9–29.0	29.8–25.1–21.3	MCMCTree	plastomes	14 fossils[c]
Mandáková et al.[39]	54.7–40.1–29.4	30.6	BEAST	plastomes	four fossils
Huang et al.[23]	33.2–29.9–26.8	22.9–21.3–19.6	BEAST	plastomes	four fossils
Ramírez-Barahona et al.[83]	52.7–41.9–30.5	–	BEAST	rbcL, atpB, matK, ndhF, 18S, 26S, 5.8S	238 fossils, angiosperm-wide
Walden et al.[2]	35.7–29.9–24.3	29.6–25.1–20.9	BEAST	plastomes	four fossils
Legalov et al.[84]	–	>36.4	indirect calibration via phytophagous beetles	–	*Ceutorhynchus* beetle fossils
This study (nuclear dataset)	25.7–24.5–23.1	22.4–21.1–19.9	treePL	phylogeny using superstrict routine; calibration using subset of 20 most clock-like genes	one fossil validated by nine secondary calibration points
This study (plastome dataset)	29.0–20.2–13.0	24.3–16.9–10.2	treePL	plastomes	one fossil validated by eight secondary calibration points

Comparison of estimates from past studies and the current study (taken and updated from Huang et al.[23]).
[a]Core Brassicaceae are all Brassicaceae, excluding basal tribe Aethionemeae.
[b]Dating results excluding *Thlaspi primaevum*.
[c]Only results that exclude Brassicales fossils.

Merci de votre attention

Des questions?

I- Reconstruction phylogénétique des Brassicacées

2) Sélection de gènes orthologues putatifs :

Méthodes de Sélection des Gènes Orthologues putatifs

- **Méthode n°1 :** Sélection de **28 gènes issus d'analyses précédentes** et validés par phylogénie monogénique dans 13 Brassicacées.

- **Méthode n°2 :** Sélection de **45 gènes** parmi **deux bases de données d'orthologues couvrant 9 Angiospermes, en vérifiant leur conservation dans les Brassicacées**.

- **Méthode n°3 :** Sélection de **40 gènes orthologues à copie unique conservés entre 5 espèces de Brassicacées qui sont également conservés chez les Angiospermes.**

113 gènes combinés issus des trois groupes, formant une base robuste pour des analyses phylogénétiques…

Méthodes de Sélection des Gènes Orthologues putatifs

Groupe de gènes	Origine des données	Critères de sélection	Résultat
Premier groupe	Études précédentes (Zhang et al. 2012 ; Zeng et al. 2014) utilisant HaMStR et OrthoMCL sur 9 génomes d'angiospermes.	- Reconstruction de phylogénies monogéniques avec RAxML.- Regroupement des espèces par lignées.- Exclusion des duplications dues aux polyploïdisations fréquentes dans les Brassicaceae.	28 gènes retenus.
Deuxième groupe	Chevauchement entre 2 bases de données : 1. 4 180 OG (HaMStR).2. 1 989 OG (OrthoMCL) chez 7 angiospermes.	- Partagés par les deux bases de données.- Couverture > 80 % des 13 espèces.- Protéines > 300 acides aminés.- Regroupement cohérent des lignées dans les arbres monogéniques.	45 gènes retenus.
Troisième groupe	Chevauchement entre :1. 4 180 OG (HaMStR).2. 6 552 OG (OrthoMCL) à copie unique dans 5 Brassicaceae.	- Copie unique conservée dans les angiospermes et Brassicaceae.- Couverture > 80 % des 13 espèces.- Protéines > 500 acides aminés.- Regroupement cohérent selon les trois lignées.	40 gènes retenus.
Total (3 groupes)	Combinaison des trois approches.	- Analyse rigoureuse des phylogénies monogéniques.- Maximisation de la qualité et de la pertinence des marqueurs génétiques retenus.	**113 marqueurs**

Méthode 1 : Gènes orthologues conservés chez les angiospermes

Objectif : Identifier des gènes nucléaires à faible nombre de copies conservés chez les angiospermes et les adapter aux Brassicacées.

Procédure :

1. Les gènes ont été récupérés à l'aide de deux outils :
   - **HaMStR** (Ebersberger et al. 2009).
   - **OrthoMCL v1.4** (Li et al. 2003).
2. Ces outils ont permis d'identifier 106 gènes partagés par neuf espèces d'angiospermes ayant un génome séquencé (*Z. mays* , *S or . bicolor* , *O. sativa* , *G. max* , *M e . truncatula* , *P o . trichocarpa* , *V. vinifera* , *A. thaliana* et *Sol. lycopersicum*).
3. Les homologues de ces gènes ont ensuite été récupérés dans les génomes de 13 espèces de Brassicacées.
4. Chaque gène a été évalué à l'aide de phylogénies monogéniques (avec RAxML), en vérifiant que les espèces de la même lignée (LI, LII, ou LIII) étaient correctement regroupées, sans exiger des relations spécifiques entre ou au sein des lignées.

Résultat : 28 gènes ont été retenus comme marqueurs orthologues fiables.

Méthode 2 : Orthogroupes à faible nombre de copies partagés par les angiospermes
Objectif : Élaborer un deuxième ensemble de marqueurs en utilisant des bases de données indépendantes pour diversifier la sélection.
Procédure :

1. **Deux bases de données ont été combinées :**

   - Une contenant 4 180 groupes de gènes nucléaires orthologues (OG) identifiés par HaMStR à partir de neuf espèces d'angiospermes.

   - Une autre avec 1 989 OG à faible nombre de copies identifiés par OrthoMCL à partir de sept espèces d'angiospermes (*A. thaliana*, *P. trichocarpa*, etc.).

2. Les OG partagés par ces deux bases de données ont été considérés comme conservés et à faible nombre de copies.

3. Critères supplémentaires pour sélectionner des gènes fiables :

   - Présence dans au moins 80 % des 13 espèces de Brassicacées.

   - Codage de protéines de plus de 300 acides aminés.

4. Les arbres monogéniques de ces gènes ont été analysés pour vérifier que les espèces se regroupent correctement selon les trois lignées des Brassicacées.

Résultat : 45 gènes ont été retenus pour ce deuxième groupe.

Méthode 3 : Gènes conservés chez les angiospermes et les Brassicacées

Objectif : Identifier des orthologues à copie unique dans les Brassicacées, qui sont également conservés chez les angiospermes.

Procédure :

1. Les orthologues ont été extraits de deux ensembles de données :
   - Les 4 180 OG identifiés par HaMStR (voir Méthode 2).
   - Les 6 552 OG identifiés par OrthoMCL comme des orthologues à copie unique dans cinq Brassicacées (*A. lyrata*, *B. rapa*, etc.).
2. Critères supplémentaires pour sélectionner des gènes fiables :
   - Présence dans plus de 80 % des 13 espèces représentatives de Brassicacées.
   - Codage de protéines de plus de 500 acides aminés.
3. Les arbres monogéniques ont été vérifiés pour garantir que les espèces se regroupent correctement selon les trois lignées des Brassicacées.

Résultat : 40 gènes ont été retenus pour ce troisième groupe.

1. Préparation des séquences

- **Obtention des gènes** :

Les 113 gènes orthologues ont été récupérés à l'aide de **HaMStR**, avec une valeur seuil d'E < e-20 (gènes significativement similaires).

- **Alignement** :

Les séquences nucléotidiques ont été alignées en utilisant **MUSCLE** (un logiciel d'alignement de séquences) avec ses paramètres par défaut.

- **Nettoyage des alignements** :
  - Les séquences de mauvaise qualité ont été supprimées.
  - Les régions mal alignées ont été réduites à l'aide de **trimAl**, garantissant des alignements fiables.
- **Concaténation** :

Les alignements des gènes individuels ont été concaténés pour former une seule matrice à l'aide de **SeaView**.

- **Gestion des duplications** :

Pour les gènes avec plusieurs copies dans une espèce, la copie avec la branche la plus courte (la plus proche de l'orthologie supposée) a été conservée pour éviter les paralogues.

a) Inférence bayésienne (BI) - MrBayes :

- **Paramètres de l'analyse** :
  - Utilisation de **MrBayes** avec 5 000 000 générations.
  - Application de la méthode **MCMC (Markov Chain Monte Carlo)** pour estimer la probabilité des arbres :
    - 1 chaîne froide et 3 chaînes chaudes pour explorer l'espace des arbres phylogénétiques.
  - Un seuil d'arrêt automatique (**stopval = 0,01**) pour garantir la convergence.
- **Suivi de la convergence** :

La convergence des chaînes a été contrôlée avec **AWTY**, un outil permettant de visualiser si les chaînes ont atteint une solution stable.

- **Échantillonnage des arbres** :
  - Les arbres ont été échantillonnés toutes les 100 générations.
  - Les 25 % premiers arbres (phase de burn-in) ont été exclus.
  - Les arbres restants ont été utilisés pour générer un **arbre de consensus**.

b) Maximum de vraisemblance (ML) - RAxML :

- **Paramètres de l'analyse** :
  - Les analyses ont été réalisées avec **RAxML**, initialement avec le modèle **GTRGAMMA**, mais la topologie et les valeurs bootstrap étaient équivoques.
  - Pour optimiser l'efficacité, le modèle **GTRCAT** (approximation plus rapide de GTRGAMMA) a été utilisé pour des analyses supplémentaires.
- **Bootstrap** :

Les scores de bootstrap (évaluant la robustesse des branches) ont été calculés, bien que les résultats initiaux aient montré des ambiguïtés.