Génomique évolutive de familles de gènes

Nathalie Chantret

UNIVERSITE MONTPELLIER II

ECOLE DOCTORALE GAIA

DOSSIER DE CANDIDATURE

HABILITATION A DIRIGER DES RECHERCHES

Nathalie Chantret

Génomique évolutive de familles de gènes

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d’HDR, les valeurs et principes d’intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l’article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d’intégrité scientifique de l’Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

SOMMAIRE

Avant-propos

Si je devais décrire en quelques mots ce qui m'a menée au métier de chercheur, je dirais qu'initialement ce fut l'intérêt pour les sciences naturelles, le plaisir de découvrir et de comprendre comment les choses fonctionnent, d'apprendre, et aussi un certain émerveillement esthétique devant la nature. La liste de ces motivations est longue et a évolué au cours du temps, elle s'est enrichie aussi, en particulier du gout du travail collaboratif. Il y a aussi l'attrait pour la 'démarche scientifique'. Malgré un certain consensus sur cette démarche, elle renferme deux volets qui semblent un peu opposés mais sont, à mon sens, assez complémentaires. Le premier est l'aspect exploratoire, qui inclut la curiosité, le gout pour sortir des sentiers battus, pour fouiller, parfois à perte, dans la masse des informations, prospecter et remettre parfois en question les 'dogmes' et rester ouverts à des idées originales voire 'tordues'. Ces aspects tiennent de la créativité, du désordre ou de l'anticonformisme. Le deuxième volet est l'aspect purement rationnel, celui qui conduit à appliquer des méthodes précises, calibrées, faire des tests extrêmement nombreux et répétés qui décomposent les facteurs et qui, pour être menés correctement et permettre des déductions précises, nécessitent l'assimilation d'un corpus de connaissance d'une grande complexité, souvent le fruit de décennies de constructions expérimentales et théoriques. Même si le premier aspect correspond plus au 'cliché' du chercheur, c'est sur le deuxième que repose l'essentiel de notre travail, et, bien sûr, les hypothèses que le premier ferait émerger doivent se confronter aux méthodes du deuxième pour être validées ou tomber dans l'oubli, au moins un certain temps. Il est idéal que ces deux aspects soient présents, au sein d'une communauté scientifique mais aussi, dans une moindre mesure, chez chacun, sachant que l'équilibre entre les deux est variable et très personnel.

Une autre façon de parler de ce métier serait aussi d'en examiner les composantes sur un gradient entre raison et émotion au sein des hommes et des femmes qui l'exercent et lors de leurs échanges ; à nouveau, de façon simpliste, on peut décrire une des extrémités de ce gradient comme étant de la rationalité pure, qui peut s'exercer individuellement, voir dans l'isolement. Cette vision théorique reste une vue de l'esprit parce que les acteurs sont des humains et les humains ne peuvent pas mettre leurs émotions au placard quand ils commencent leur journée de travail. Et surtout, en plus d'être difficile, les mettre au placard serait une erreur, de mon point de vue. Les très rares personnes que j'ai pu croiser, qui tentent de laisser leurs émotions au placard, voire qui y parviennent, génèrent souvent autour d'eux des malaises, du mal-être, jusqu'à des angoisses, chez leurs collègues. Éteindre ses émotions mène souvent à éteindre son empathie, ce qui peut conduire à traiter de façon incorrecte les individus. Ce n'est manifestement pas ce à quoi nous aspirons dans notre milieu professionnel. La science n'a pas une valeur intrinsèque plus haute que les personnes, elle ne passe pas 'avant' les personnes. L'inverse n'est pas non plus vrai. La science n'existe et n'a de sens qu'avec et pour les personnes. À nouveau c'est une question d'équilibre. Le véritable défi est de mener des recherches de qualité dans le respect de l'ensemble des individus qui les mènent.

Il ne s'agit pas d'insuffler et de mixer des 'sentiments humains' dans la science, mais juste de se rappeler que les émotions des personnes qui l'exercent sont utiles et vitales. Elles sont utiles pour les personnes elles-mêmes - qui va bien travaille bien - et elles sont utiles pour l'émulation intellectuelle ; l'homme est social comme on dit, il échange et il confronte ; exposer à autrui, communiquer pour se faire comprendre, formuler, utiliser le verbe, est un processus d'une grande efficacité et dont il est totalement absurde de se priver.

Aujourd'hui, la culture de la performance met quand même beaucoup en exergue les qualités individuelles, la combativité ou la pugnacité individuelle, et ne promeut ou ne valorise peut-être pas encore assez le travail collaboratif réel au quotidien ou la diversité et la complémentarité des personnes. Plus récemment, la distanciation sociale due aux méthodes de lutte contre la pandémie a eu des conséquences évidentes sur les interactions scientifiques entre individus. Cette période aura eu au moins l'avantage de nous rappeler, à tous, à quel point ces échanges du quotidien sont importants - et ils sont essentiels pour moi.

Mon parcours en quelques mots

Après deux années de classes préparatoires à Paris (lycée Henry IV) j'ai intégré l'École Nationale Supérieure Agronomique de Rennes (ENSAR, maintenant Institut Agro Rennes-Angers) en 1992 et choisi la spécialité 'Amélioration des Plantes', en troisième année (dirigée à l'époque par le Professeur Yves Hervé). Cette dernière année, j'ai combiné un DAA (Diplôme d'Agronomie Approfondie) et un DEA (Diplôme d'Études Approfondies), délivrés par l'école agronomique, afin de pouvoir poursuivre mes études en réalisant un doctorat. J'avais réalisé mon stage de master dans le laboratoire de l'INRA de Rennes-Le Rheu sur la caractérisation, par des méthodes cytogénétiques, d'une introgression portant des gènes de résistance à un virus chez le blé tendre avec Joseph Jahier. À l'issue de cette troisième année j'ai été reçue au concours 'ASC' (Agent Scientifique Contractuel) de l'INRA (département DGAP) pour effectuer ma thèse à la station d'Amélioration des Plantes de l'INRA de Rennes sous la direction de Gérard Doussinault.

Mon sujet de thèse était « Étude génétique et marquage moléculaire des facteurs de résistance à l'oïdium du géniteur de blé tendre RE714 ». Outre son rôle de formation à la recherche, cette thèse m'a permis d'aborder plusieurs notions, et d'acquérir un ensemble de connaissances, listées ci-après sans hiérarchie : marquage moléculaire et cartographie génétique, pathologie végétale et protocoles expérimentaux de tests de résistance (en conditions contrôlées et en champ), génétique quantitative appliquée et recherche de QTL, cytogénétique, connaissance du blé et des céréales (génétique et polyploïdie, croisements, fixation de matériel). Au-delà des aspects scientifiques, cette thèse a eu un rôle très important dans mon apprentissage de la recherche au sens large, en particulier celui d'apprendre à remettre en question certains résultats, à écouter ses doutes et à creuser certaines pistes quitte à les abandonner après, ou à travailler en équipe et rédiger des articles scientifiques.

À la suite de cette thèse, animée par l'envie d'ouvrir mes horizons en termes d'expériences scientifiques et humaines, je suis partie en stage postdoctoral aux USA en 1999, dans le laboratoire de Jorge Dubcovsky, Université de Davis, Californie, et j'ai basculé dans le monde de la génomique. Mon projet était alors d'élucider, d'un point de vue génomique, les causes de l'absence des gènes responsables du caractère 'tendre' dans le blé dur. Le grain de blé tendre, espèce *Tricicum aestivum*, a un endosperme friable, ou tendre, parce qu'il possède ces gènes qui confèrent ce caractère de tendreté. Le grain de blé dur, espèce *Triticum durum*, les a perdus et son endosperme, beaucoup plus difficile à fragmenter, est qualifié de dur. Cette différence majeure entre ces deux espèces est particulièrement importante pour leur utilisation en alimentation (blé tendre facilement réduit en farine et blé dur en semoule). Pour atteindre ce but, l'étape majeure qui a occupé la majorité de mon temps de post-doctorat, fut de construire une banque BAC de blé dur (Partie I). Lors de ce séjour, en plus des compétences de biologie moléculaire, j'ai appris à manipuler les séquences de grands fragments et fait mes premières expériences de bio-analyse. Ce séjour outre-atlantique fut, là encore, extrêmement enrichissant d'un point de vue scientifique, mais aussi culturel et humain.

En septembre 2001 j'ai été recrutée sur un poste de chercheur INRA positionné au CIRAD dans l'UMR GACA (Génomique Appliquée aux Caractères Agronomiques) devenue PIA (Polymorphismes d'Intérêt Agronomiques) un peu plus tard. J'ai rejoint l'équipe 'évolution des génomes', valorisé mes travaux de post-doctorat et participé à un certain nombre de projets de génomique. En 2006 j'ai rejoint l'UMR 'DIA-PC' (Diversité et Adaptation des Plantes Cultivées) et l'équipe travaillant sur la dynamique de la diversité, notamment le groupe travaillant sur l'espèce modèle *Medicago truncatula*, sur le site INRA de Mauguio. Nous avons déménagé sur le campus de l'école d'agronomie de Montpellier en 2012, puis en 2019 dans le bâtiment 'ARCAD' tout neuf, sur le campus de Lavalette. C'est suite à cette nouvelle affectation que j'ai découvert un ensemble de champs disciplinaires, incluant notamment la génétique des populations et l'évolution moléculaire et de nouveaux axes de recherche. À cette période, j'ai également commencé à travailler avec de nouveaux collègues, ceux aux côtés desquels je travaille encore actuellement, ce qui, de mon point de vue, mais j'espère qu'il est partagé, est la preuve que l'interaction fonctionne plutôt bien depuis plus de 15 ans.

CURRICULUM VITAE DETAILLE

Nathalie Chantret

Née le 21 Aout 1972 ; Nationalité française ; 3 enfants

Cursus académique

Nov. 95 - Avr. 99 — **Doctorat** de l'Ecole Nationale Supérieure d'Agronomie de Rennes (ENSAR) Soutenance le 8 avril 1999.

Sept. 94 - Sept. 95 — **DEA** et **DAA** de l'Ecole Nationale Supérieure d'Agronomie de Rennes (ENSAR).

Expériences et parcours professionnels

2005 - … — **CR1** INRAE, en 2006 changement pour Unité DAP[1], puis AGAP[2], AGAP Institut.

2001 - 2005 — **CR2** INRAE, Unité GACA[3], puis PIA[4].

Oct. 99 - Avr. 01 — **Post-doctorat**, Université de Davis, Californie, « Etude de la différenciation des génomes homéologues chez les blés dans les régions des gènes codant pour les puroindolines. Création et exploitation d'une banque BAC de blé dur. » PI Jorge Dubcovsky; Contrat ASC INRA.

Nov. 95 - Avr. 99 — **Doctorat**, INRA de Rennes-Le Rheu, « Facteurs de résistance à l'oïdium chez le géniteur de blé tendre RE714 ». Directeur Gérard Doussinault ; Contrat ASC INRA.

Nov. 95 - Sept. 01 — **Attachée Scientifique Contractuelle** INRAE, GAP.

Jan. 95 – Juill. 95 — **Stage de DEA-DAA**, INRA de Rennes-Le Rheu, « Evaluation cytogénétique de lignées de recombinaison de blé tendre résistantes à la jaunisse nanisant de l'orge ». Direction Joseph Jahier.

Encadrement d'étudiants

NB : pour plus de détails cf. rubrique 'Encadrement d'étudiants' p.80

<u>*Post-doctorat*</u>

Iris Fischer — avril 2012 - juin 2014 : financement Agropolis fondation (ARCAD project);
juill. 2014 - juin 2016 : financement 'Deutsche Forschungsgemeinschaft'

Jacques Dainat — jan. 2013 - dec. 2013 : financement Institut Agro Montpellier

[1] Diversité et Adaptation des Plantes

[2] Amélioration Génétique et Adaptation des Plantes

[3] Génomique Appliquée aux Caractères Agronomiques

[4] Polymorphismes d'Intérêt Agronomiques

Doctorat

Céline Gottin	oct. 2018 - nov. 2021 : financement CIRAD / Institut Agro Montpellier

Césure

Fabien Bustos	mars 2022 - … : Institut Agro Montpellier
Audrey Serra	juin 2015 - juil 2015 : INSA de Lyon

M2

Thibaut Vicat	fev. 2021 - juil. 2021 : M2 bioinformatique UM II
Enora Gesclin	nov. 2019 - juil.2020 : M2 bioinformatique Univ. Rennes
Asya Martirosyan	mars 2015 - sept. 2015 : M2 UMII parcours MEME[5]
Maxime De Sario	jan. 2014 - juin 2014 : M2 UMII parcours BEE[6]
Émilie Roux	jan. 2009 - juin 2009 : M2 UMII parcours DEPS[7]
Joan Ho-Huu	jan. 2008 - juin 2008 : M2 UMII parcours DEPS[7]
Saïfallah Bousselmi	fev. 2007 - juin 2007 : M2 UMII Supagro parcours RPIB[8]

M1

Fadwa El Khaddar	avril 2022-… : Master I Bioinformatique UM II
Quentin Oliveau	mai 2012 : AgroParisTech 2ème année

Participation à des jurys thèse/master

Jury Master

Jury Master II Parcours Darwin : Biologie Evolutive & Ecologie BEE 2019; **rapporteur de 4 étudiants**, examinateur pour les 11 autres.

Jury Master

Rapporteur Master II Supagro Apimet-Sepmet 2012-2013 «Diversité moléculaire au locus Pc5.1, un QTL clé de la résistance partielle du piment à *Phytophthora capsici* » **Hélène Pidon** ; encadrement Véronique Lefèbvre.

Jury de thèse

Examinateur ; **Karine Charon**, 2007 « Caractérisation fonctionnelle et évolution moléculaire des gènes codant pour les facteurs d'initiation de la traduction eIF4E : des facteurs clés dans la résistance des plantes aux potyvirus » ; GAFL INRA Montfavet ; encadrement Carole Caranta.

[5] 'Erasmus Mundus Master Programme in Evolutionary Biology'

[6] Biologie Evolutive et Ecologie

[7] Diversité et Evolution des Plantes et de leurs Symbiotes

[8] Ressources Phytogénétiques et Interactions Biologiques

Comités de thèse

Chloé Beaulieu
2021- … « Exploration des interactions entre mutualisme et parasitisme aux échelles micro et macro-évolutives » ; LRSV ‘Laboratoire de Recherche en Sciences Végétales’ Toulouse ; direction Maxime Bonhomme.

Félicien Favre
2021- … « Identification de facteurs génétiques impliqués dans la résistance à la fusariose du vanillier par des approches de génomique et génotypage haut débit » ; UMR PVBMT, La Réunion ; direction Pascale Besse et Carine Charron.

Julio de Andrade Garighan
2018-2021 « Study of the temperature-mediated transcriptional and post-transcriptional regulation of dormancy and bud break in apple tree »; UMR AGAP Montpellier équipe AFEF, Montpellier ; direction Evelyne Costes et Fernando Andres.

Camille Gréard
2016-2017 « La détection de variants alléliques comme voie d’amélioration génétique des plantes fourragères. Exemple de la luzerne » ; UMR P3F – UM Pluridisciplinaire Prairies et Plantes Fourragères, Lusigan ; direction Bernadette Julier.

Enrique Abboud-Ortega
2011-2013 « Comparative genomics on the secreted proteins of the agent of rice blast *Magnaporthe oryzae* » ; UMR BGPI Montpellier, direction Didier Tharreau et Elisabeth Fournier.

Lamis Karaki
2010-2013 « Recherche et diversité de molécules entomotoxiques de la famille des albumines A1b dans différentes espèces de Légumineuses originaires du Proche-Orient » ; UMR Biologie Fonctionnelle Insectes et Interaction ; INRA INSA-Lyon ; direction Yvan Rahbé.

Romain Philippe
2007-2008 « Diversité ortho-allélique de deux familles multigéniques (ASR et invertases vacuolaires) candidates pour la tolérance à la sécheresse chez les céréales » ; UMR PIA Montpellier ; direction ; Dominique This.

Martine Leflon
2004-2007 « Introgression et stabilisation d’un cluster de gènes de résistance spécifiques porté par un génome diploïde (Brassica rapa, AA) dans un génome allotétraploïde (B.napus, AACC) » ; UMR APBV Rennes ; encadrement Anne-Marie Chèvre et Eric Jenczewski.

Enseignement

- DESS ‘Génétique, Génomique et Technologies Avancées des Végétaux’, UM II, Responsable Pr. B. Touraine : 3h de cours et 3h de TD par an de 2002 à 2004.

- DEA ‘Développement et Adaptation des Plantes’ ; module "analyse et exploitation des génomes végétaux", UM II, Responsable Pr. M. Lebrun : 3h de cours par an de 2002 à 2004.

- Master 2 ‘Ressources Phytogénétiques et Interaction’ UM II : 2h de cours 2006.

- Master 2 ‘DEPS’ Supagro (resp. Pr. J. David) ; 3h de cours par an de 2007 à 2009.

- Formation continue Supagro Montpellier bioinformatique (resp. D This) Intervention sur le thème de ‘recherche de séquences et analyses phylogénétiques’. 2011 : 8 heures.

- Formation ARCAD (SP1-SP4) ‘Analyse des données de polymorphisme’ Montpellier. Intervention ‘Annotations simples et gestion des familles de gènes via les outils de phylogénie : identification d’orthologues’ 2011 : 8 heures.

Projets et financements

Nb : les projets en gris clair italique ont été soumis mais pas retenus

*- 2021 ‘**RIR** Rice Immune Receptors: Diversity and evolutionary dynamic of rice LRR immune receptors’ **Agropolis fondation Open Science**; co-porteuse avec Thomas Kroj (PHIM).*

- **2019-2020 -** ‘**CAS** Conservation of Alternative Splicing’ ; **Dépt. INRAE BAP**; 28k€ ; *porteuse*

*- 2019 - ‘**DHEA** Diversity, Heritability, and Evolution of Alternative Splicing in Plants’ **ANR** - porteuse*

*- 2018 - ‘**INTERPRELCO** Deciphering the molecular network of cell surface immune receptors for resistance to pathogens in crops by in silico INTERaction PREdictions within Lrr-COntaining proteins’ **ANR** Appel d’offre ‘SusCrop- ERA-NET’; porteuse Anne Diévart.*

- **2016-2020 -** ‘**REGULEG** Identifying regulators of legume seed adaptation to environmental changes’ (projet générique); **ANR** ; porteuse Julia Buitnik, Anger; 562 k€ ; *partenaire : implication dans 2 WP*.

*- **2016** – ‘ALTERNATIVE Alternative splicing and adaptation: plant domestication as a case study’ **Agropolis fondation Open Science**; co-porteuse avec Christopher Sauvage (GAFL).*

- **2014-2015 -** ‘**DEFENSIN** Tackling the DEFENSIN family Evolution within the extremophile species *Arabidopsis halleri’* ; **Projet Agropolis Fondation** Open Science – Reseach; porteuse Françoise Gosti, UMR BPMP; 28 k€ ; *partenaire : implication dans 3 WP*.

- **2012-2016** - ‘**TRANS** Breeding system transitions in flowering plants and their genomic consequences’; **ANR** ; porteur Sylvain Glémin, ISEM, Montpellier; 532 k€ ; *correspondante de l’équipe et porteuse du WP10* ‘Breeding systems and the fate of duplicated genes’.

- **2011-2014 -** ‘**ARCAD SP1** Comparative population genomics in wild and crop plants: a genome and phylogenetic wide approach’; **Projet Agropolis Fondation** porteurs Pr. Jacques David, Supagro Montpellier et Sylvain Glémin, ISEM, Montpellier ; 1.100 k€ ; ARCAD ‘Agropolis Resource Center for Crop Conservation, Adaptation and Diversity’ est un projet Etendard Agropolis Fondation (porteur Jean-Louis Pham, IRD, Montpellier) ; *co-porteuse du WP5* ‘Comparative functional genomics’ *et référente espèce Medicago sativa* (WP1).

- **2011-2015 -** ‘**IMMUNIT-AE** Genetic diversity and mechanisms of resistance to *Aphanomyces euteiches* in legumes’; **ANR**; porteur Pr. Christophe Jacquet, Univ. Paul Sabatier, Toulouse III ; 700 k€ ; *correspondante de l’équipe et implication dans 2 WP*.

Activités d'expertise

Evaluation d'articles scientifiques de type 'pair-reviewing' :

Plus de 20 articles expertisés pour les journaux à comité de lecture : TAG (3), Genome (1), Mol Genet & Genomics (1), Genetica (1), BMC Plant Biol (5), BMC Genomics (3), Plant Mol Biol (3), BMC Evol Biol (1), Annals of Botany (1), New Phytol (1), Heredity (1), FEBS journal (1) …

Expertise de projets :

- Sujet de thèse BAP (2011) et EFPA (2017)
- Projet soumis à l'Appel d'Offre INRA/RUSSIE (2009).
- Projet soumis à la National Science Fundation (2008)

Autres activités collectives

- Animation du groupe 'bio-informatique' formé en 2020 (séminaires mensuels).

- Membre du **comité d'organisation des journées d'inauguration du bâtiment ARCAD** grand public, conjointement à la fête de la science (organisation, préparation de supports type plaquette, poster, participation aux journées portes-ouvertes avec animation d'un atelier 'biologie moléculaire : ADN-ARN-PCR le secret de la vie') Octobre 2021.

- Depuis 2011 membre des **encadrants du groupe des doctorants INRAE-BAP dit « Jeunes chercheurs ».** NB : Cette tâche consiste à assister/animer aux réunions du groupe (une à deux fois par an, deux jours). Les premières années, il s'agissait aussi la relecture et l'accompagnement des étudiants dans la rédaction de leur résumé de thèse pour constituer le cahier des résumés des doctorants, produit une fois par an, qui n'est plus produit depuis que le département a évolué.

- Participation à l'organisation des rencontres du groupe inter-institut 'cytogénétique-polyploïdie' qui a eu lieu à Montpellier en 2004.

- Chercheur référent de Karine Loridon Ingénieur dans l'équipe de 2010 à 2019 et d'Isabelle Hochu (2008-2009).

- Participation aux journées d'animation scientifique « Biologie Intégrative de l'Adaptation des Plantes à l'Environnement (BIAPE) » organisées par le département BAP les Juin 2013.

- Formation de techniciens : Audrey Weber et Isabelle Hochu à l'utilisation de logiciels de génomique (éditeur de séquence) et à l'alignement de séquence 2006.

- Correspondante du département GAP pour l'UMR PIA (2003- 2005)

- Organisation du déménagement de la banque BAC de blé dur (réalisée pendant mon stage postdoctoral et achetée par l'INRA) de Montpellier au Centre National de Ressources Génomiques Végétales (INRA – Toulouse)

Liste résumée des productions scientifiques

NB : pour plus de détails cf. liste des productions scientifiques

(a) **30 articles publiés dans des revues à comité de lecture**.

Tableaux présentant quelques métriques de publication : nombres de papiers par date de publication, par nombre de citations et par facteur d'impact du journal (réalisé en avril 2022).

dates de publication	nb. papiers
2018 - 2021	6
2014 - 2017	8
2010 - 2013	3
2006 - 2009	4
2002 - 2005	6
avant 2002	3

nb. citations	nb. papiers*
> 100	4
50 - 100	7
20 - 50	8
10 - 20	6
< 10	5

IF journal**	nb. papiers
>15	2
10 - 15	4
5 - 9	9
3 - 4	10
< 3	4

* **h-index 19** : 19 articles cités au moins 19 fois ('Hirsch Index').

** IF : Facteur d'Impact du journal

(b) **Un chapitre d'ouvrage**

Ranwez, V., and N. Chantret. **2020**. 'Strengths and Limits of Multiple Sequence Alignment and Filtering Methods.' in C. Scornavacca, Delsuc, F., and Galtier, N., editors (ed.), Phylogenetics in the Genomic Era.

(c) **16 posters et 9 communications orales**

Collaborations productives

A partir de 2006

Au sein de mon équipe GE^{2}pop:

- Ardisson Morganne
- Burgarella Concetta (en post-doctorat)
- David Jacques
- Freville Hélène
- Gay Laurène
- Latreille Muriel
- Loridon Karine
- Muller Marie-Hélène
- Prosperi Jean-Marie
- Ronfort Joëlle
- Santoni Sylvain
- Ranwez Vincent
- Tavaud-Pirra Muriel
- Tollon-Cordet Christine
- Weber Audrey

Au sein de mon unité AGAP :

- Diévart Anne
- Droc Gaétan
- Dufayard Jean-François
- Gautier Marie-Françoise
- Perin Christophe
- Pot David
- Sarah Gautier
- Summo Marilyne

Autres unités à Montpellier :

- Cenci Alberto (Bioversity International)
- De Mita Stéphane (PHIM)
- Glémin Sylvain (ISEM)
- Kroj Thomas (PHIM)

Autres unités en France :

- Bonhomme Maxime (LRSV Toulouse)
- Buitink Julia (IRHS Angers)
- Jacquet Christophe (LRSV Toulouse)
- Pilet Marie-Laure (IGEPP Rennes)
- Plomion Christophe (BIOGECO Bordeaux)
- Rahbé Yvan (BF2I Lyon)

Hors de France:

- Bataillon Thomas (BiRC, Aarhus, Denmark)
- Young Nevin (University of Minnesota, St. Paul, USA)

SYNTHESE DES TRAVAUX

Les thématiques de recherche et les disciplines que j'ai abordées ont évolué au cours du temps. De façon schématique, comme évoqué dans l'avant-propos, après mon doctorat, je me suis tout d'abord orientée vers la **génomique structurale**, à une période où le rythme des découvertes sur les génomes complexes était soutenu. Puis, il y a une quinzaine d'années, j'ai élargi mes compétences en abordant de nouvelles disciplines notamment **l'évolution moléculaire** et la **génomique des populations**. J'ai pu utiliser un vaste panel de méthodes de bio-analyse et de bio-informatique, dédiées à l'analyses de séquence (alignement, phylogénie, modèles d'évolution moléculaire), devenues incontournables pour traiter l'avalanche de données générées par le séquençage de type haut débit.

Dans la synthèse de mes travaux je vais présenter le cheminement qui m'a conduit à aborder aujourd'hui des questions de recherche qui peuvent être résumées de la façon suivante : **étude des mécanismes moléculaires ou évolutifs connus dans les génomes pour documenter leur variabilité aux échelles intra- et interspécifiques et déterminer dans quelle mesure ils peuvent être le support de l'adaptation**. La première partie (I) décrit ma contribution à la **compréhension de l'organisation des génomes des plantes et de leur évolution structurale**. La deuxième partie (II) décrit mon **immersion dans la génomique des populations et l'évolution moléculaire**, et comment j'ai pu mobiliser ces nouvelles approches pour articuler des questions autour de **l'évolution des gènes dupliqués**. La troisième partie (III) est consacrée au volet peut-être le plus central de mes recherches et qui porte sur **l'étude des familles multigéniques et leur rôle potentiel dans l'adaptation**. J'ai analysé l'évolution de plusieurs familles de gènes de fonctions connues, ou sans a priori fonctionnel, à plusieurs échelles phylogénétiques. Enfin la dernière partie (IV) est consacrée à un thème que j'ai abordé plus récemment, **l'évolution de l'épissage alternatif et son rôle potentiel dans l'adaptation**.

*Les citations de type [**Pxx**] font références à mes publications listées dans la rubrique 'Liste des productions scientifiques'.*

I. Génomique structurale comparative chez les blés et leurs espèces apparentées

Les connaissances dans le domaine de la génomique ont énormément évolué ces vingt dernières années. J'ai eu la chance de pouvoir suivre en direct différentes étapes de cette évolution, et y participer. Ces avancées se sont accélérées plus récemment avec l'évolution technologique des méthodes de séquençages.

Les années de recherche que j'ai effectuées dans le domaine de la **génomique structurale comparative** m'ont profondément marquée. Cette discipline vise à décrire les génomes, les comparer et émettre des hypothèses sur les évènements et les facteurs qui les ont façonnés, sur l'ordre dans lequel ces évènements se sont déroulés, et si possible sur leurs bases moléculaires. Ces années ont constitué le socle sur lequel ont germé les questionnements scientifiques de la deuxième partie de ma carrière. Dans cette première partie, je ne vais pas reporter l'ensemble de mes travaux mais plutôt insister sur quelques éléments ou résultats marquants au regard de l'état des connaissances à cette période, et en fonction de l'investissement que j'ai pu y consacrer. Ces travaux correspondent à mon stage postdoctoral et aux premières années après mon recrutement à l'INRA de Montpellier, au sein de l'UMR GACA[9], devenue PIA[10] (2001 à 2006).

I.1. Contexte général : petite histoire de la génomique structurale et comparative.

Avant le développement des gros projets de séquençage des génomes complets - et bien avant la démocratisation de l'accès au séquençage massif que nous connaissons actuellement - les outils utilisés pour comparer les génomes étaient principalement les marqueurs génétiques et les cartes génétiques qu'ils permettent de construire. Pendant les années 1990, les comparaisons entre cartes permises par le transfert de certains marqueurs (type RFLP ou microsatellites) avaient ainsi mis en évidence l'existence d'une colinéarité globale (conservation globale des gènes et de leur ordre ou 'synthénie') entre les génomes, en particulier ceux des céréales, menant à la célèbre représentation en cercle concentriques de Gale et Devos (Gale and Devos 1998) (**Fig. 1**).

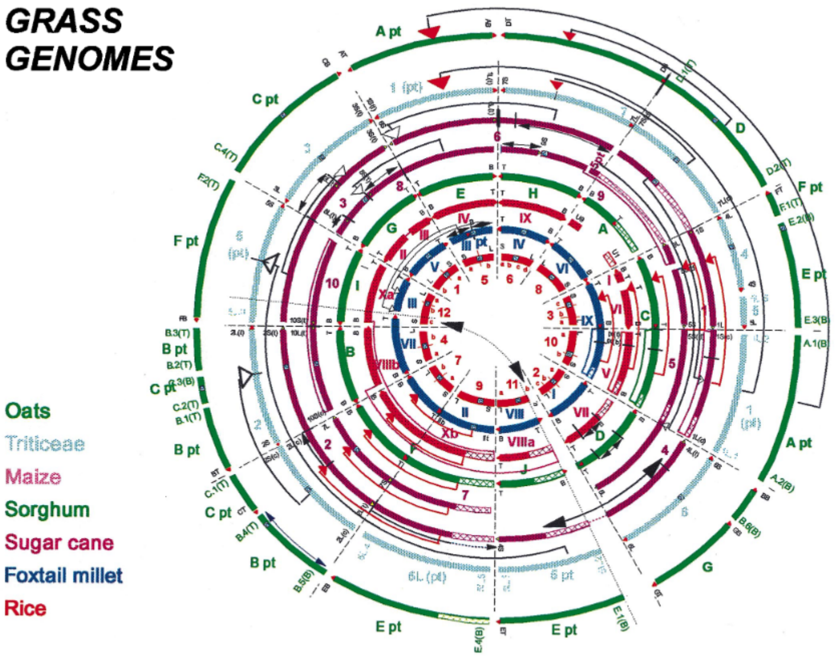


Figure 1 Colinéarité entre les génomes des céréales. Figure issue de Gale and Devos 1998 (Gale and Devos 1998) « A consensus grass comparative map drawn from many sources ».

Il s'est avéré que chez les plantes, les génomes d'espèces phylogénétiquement assez proches se ressemblent, exhibant des niveaux de colinéarité plus ou moins forts, mais présentent néanmoins des tailles très différentes. Ainsi à la fin des années 1990, plusieurs questions taraudaient la communauté scientifique des génomiciens travaillant sur les plantes : « **pourquoi les génomes des plantes ont-ils des tailles si différentes ?** », « **où sont les gènes ?** », « **à quoi correspond cette grande quantité d'ADN non codant ?** », « **quels évènements sont à l'origine de la structure actuelle des génomes ?** ».

[9] 'Génomique Appliquée aux Caractères Agronomiques'

[10] 'Polymorphismes d'Intérêt Agronomiques'

Que ce soit à des fins purement cognitives ou pour pouvoir concevoir à terme de nouvelles approches en sélection, cette communauté s'est attelée à la tâche. Une stratégie fut de choisir le chemin du 'tout séquencer', comme sur l'humain. Chez les plantes c'est l'espèce '*Arbidopsis thaliana*' qui a été choisie comme plante modèle pour plusieurs raisons, notamment la petite taille de son génome (135 Mb). Ainsi, les années 2000 marquent le début des gros projets de séquençage des génomes végétaux avec l'entreprise du séquençage du génome complet d'*Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000). La méthode utilisée à l'époque était celle des **banques d'ADN de grands fragments**, notamment de type BAC ('Bacterial Artifical Chromosome'). L'ADN génomique de la plante cible était fragmenté en grands morceaux (100 à 200 kb) puis cloné dans ces BAC. Une carte physique (de type restriction) était faite pour chaque fragment et ces cartes étaient comparées pour ordonner les BAC dans l'ordre du génome. On choisissait alors les BACs de sorte à couvrir la plus grande fraction possible du génome, avec le minimum de BAC, c'est-à-dire des BACs qui se chevauchent le moins possible ('minimum tilling path'). Ces fragments choisis étaient alors eux-mêmes fragmentés à nouveau individuellement, séquencés puis assemblés pour reconstruire la séquence de chaque BAC. C'est un travail de titan, aussi bien d'un point de vue expérimental que bioinformatique. Il a fallu plusieurs années pour arriver au bout du petit génome de cette espèce modèle.

Parallèlement, les recherches sur la famille des graminées étaient très actives. L'intérêt porté à cette famille tient au fait qu'elle contient un grand nombre d'espèces essentielles dans l'alimentation humaine à l'échelle mondiale comme le riz, le blé, le maïs, l'orge, le sorgho, le millet ou l'avoine. L'origine de cette famille a été datée entre 45 et 60 millions d'années (International Brachypodium 2010). Les génomes de ces espèces sont plutôt de grande taille. Pour donner quelques exemples la taille du génome du riz est estimée autour de 430 Mb, celle du blé tendre à environ 16000 Mb, alors que celle d'Arabidopsis est de 135 Mb. Le rapport de taille entre le blé tendre et Arabidopsis est donc de 128 et si le génome d'Arabidopsis sous forme de banque BAC tient dans une cinquantaine de plaques de 384 puits, pour stocker une banque BAC de blé dur il faut réellement un congélateur -80 entier [**P5**]. Dans la série des génomes 'difficiles' il faut citer également la canne à sucre, qui a l'originalité de présenter un nombre de chromosome variable d'un individu à l'autre (D'Hont et al. 1996). Les scientifiques travaillant sur les céréales cultivées ont donc été contraints de commencer autrement du fait de la taille beaucoup plus importante des génomes de ces espèces et ce n'est que plus tard qu'ils ont pu envisager d'appliquer la même stratégie. Pour faire avancer les recherches sur ces espèces malgré ces contraintes, c'est la mise en évidence d'un fort niveau de colinéarité entre les génomes qui a motivé la communauté à choisir une espèce modèle pour les monocotylédones, sur laquelle concentrer les efforts, en l'occurrence le riz (Feuillet and Keller 2002).

A cette période, la **génomique comparative au sein de ces espèces à grands génomes s'est donc focalisée sur certaines régions d'intérêt**. C'est dans ce contexte que s'insèrent mes travaux de post-doc et mes recherches dans les premières années après mon recrutement.

I.2. Elucider l'évolution du locus Ha dans les blés

I.2.1. Contexte et objectifs

L'objectif du projet était de **caractériser, d'un point de vue moléculaire, les changements structuraux ayant eu lieu dans une région d'intérêt du génome des blés, la région Ha impliquée dans le contrôle de la dureté du grain, et d'élucider les événements évolutifs ayant conduit à ces structures**.

Le locus Ha (pour 'hardness' : dureté) contient trois gènes paralogues (*Pina*, *Pinb* et *Gsp-1*) codant respectivement pour les protéines puroindolines a et b et pour la GSP-1 ('Grain Softness Protein'). Ces protéines présentent de fortes similarités de séquence et interviennent dans la dureté du grain de blé (démontré pour les puroindolines), caractère qui influence l'utilisation technologique finale du blé. Le complexe d'espèce des blés comporte plusieurs espèces sauvages et cultivées, et qui ont différents niveaux de ploïdie (**Fig. 2**). Les deux principales espèces cultivées sont le blé dur (tétraploïde de génomes AABB) et le blé tendre (hexaploïde de génome AABBDD). Les gènes *Pina* et *Pinb* ont été mis en évidence chez les ancêtres diploïdes donneurs des génomes A (*T. urartu*), B (*Ae. speltoides*; SS ≈ BB) et D (*T. tauschii*), ainsi que chez d'autres espèces diploïdes comme *T. monococcum* (génome A^mA^m). En revanche, ils n'ont pas pu être mis en évidence chez les blés durs tétraploïdes (de génome AABB). Chez le blé tendre issu de l'hybridation interspécifique entre un tétraploïde AABB (probablement *T. t.* ssp. *dicoccum*) et *Ae. tauschii* (de génome D), les gènes *Pina* et *Pinb* ont été réintroduit par l'intermédiaire du génome D. Ces résultats suggèrent que ces gènes, présents et fonctionnels chez les ancêtres donneurs des génomes A et B, ont été éliminés ou modifiés à la suite des événements de polyploïdisation unissant les génomes A et B pour constituer les blés tétraploïdes.

Pour étudier cette région, je me suis appuyée sur quatre espèces appartenant à la tribu des Triticées dont font partie les différentes espèces de blé (cultivées et sauvages). Ces espèces étaient (1) *Triticum monococcum*, diploïde de génome A^mA^m, (2) *Triticum turgidum* ssp. *durum*, tétraploïde de génomes AABB, le blé dur, (3) *Triticum aestivum*, hexaploïde de génome AABBDD, le blé tendre, et (4) *Aegilops tauschii*, diploïde sauvage de génome DD. Les génomes diploïdes de ces espèces ont divergé relativement récemment (**Fig. 2**) et sont donc dits 'homéologues' lorsqu'ils sont considérés au sein des espèces polyploïdes. Différents niveaux de ploïdie sont représentés dans ce complexe d'espèces, constituant ainsi un modèle particulièrement adapté pour étudier les processus évolutifs qui ont conduit à leur organisation actuelle, y compris les conséquences de la polyploïdisation.

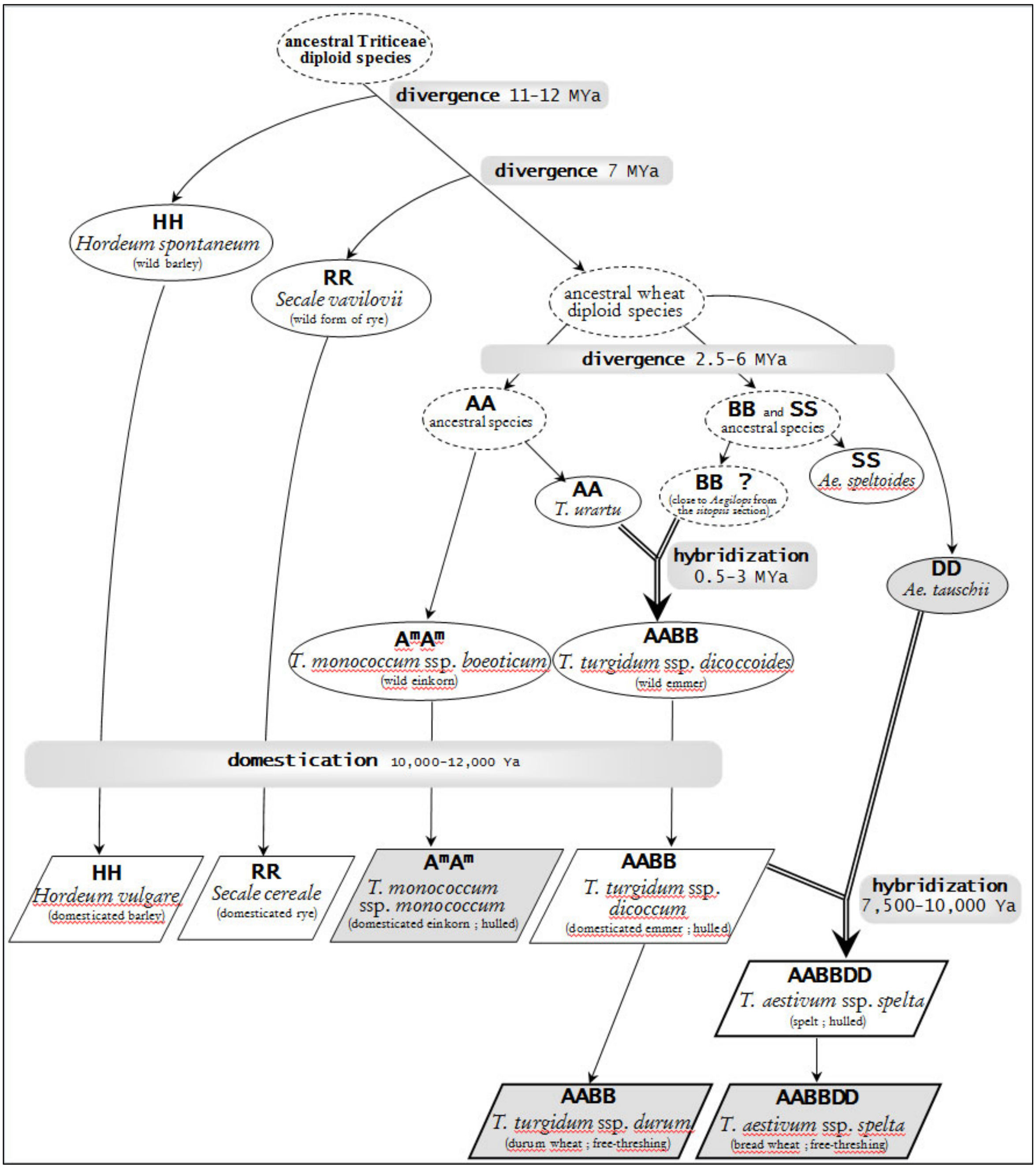


Figure 2 Représentation schématique de l'histoire des différentes espèces de blé (genres *Triticum* et *Aegilops*). Les espèces sauvages et domestiquées sont représentées respectivement dans des cercles et des rectangles. Les espèces ancestrales ou inconnues sont représentées dans des cercles en pointillé. Les espèces pour lesquelles le locus Ha a été séquencé sont grisées. Adaptée de [**P8**].

I.2.2. Obtention des séquences des sept génomes

La première espèce qui a été ciblée est *Triticum monococcum*. Une banque BAC de cette espèce avait été construite dans le laboratoire de Jorge Dubcovsky (Lijavetzky et al. 1999) et criblée avec des sondes correspondant aux gènes *Gsp-1* et *Pina*, permettant d'isoler un clone BAC contenant les trois gènes (Gautier et al. 2000). J'ai entrepris le séquençage de ce clone BAC (banque de sous-clonage, séquençage, assemblage) puis son annotation. Sept gènes ont été trouvés, en plus des gènes *Pina*, *Pinb* et *Gsp-1*. La comparaison avec le génome du riz a permis de mettre en évidence une région synténique contenant les homologues de trois autres gènes du locus, dans le même ordre et dans la même orientation. Nous avons également montré qu'un ancêtre du gène de la *Gsp-1* a pu exister avant la séparation entre les branches phylogénétiques ayant donné les riz et les Triticeae [**P7**].

Pour obtenir les séquences de *Triticum turgidum* ssp. *durum*, il a d'abord fallu **construire une banque BAC de blé dur** [**P5**]**,** une tâche à laquelle je me suis attelée et qui a constitué le socle sur lequel s'est appuyé le reste de mon projet de post-doctorat. Dans le contexte précédemment décrit, construire cette banque BAC représentait un réel défi, tant du point de vue technique (production des clones) que pour l'ampleur de la tâche car les moyens expérimentaux n'étaient pas du tout robotisés au début du projet. En plus de la pratique de la génomique d'espèces à gros génome, ce projet m'a permis de collaborer avec un laboratoire de l'USDA à l'Université de Berkeley (responsable Olin Anderson), disposant d'un robot 'Q-bot'. Organiser le transport de cette masse de clones bactériens (organismes vivants) des Etats-Unis à la France, réfrigérés à -80°C, fut un vrai chalenge (administratif et logistique).

Une banque BAC de blé tendre avait été créée à l'URGV dans le cadre de Génoplante par l'équipe de Boulos Chalhoub. De même, le criblage de cette banque a permis d'isoler 3 clones (un par génome) qui ont été séquencés.

Enfin, la banque BAC d'*Aegilops tauschii* avait été réalisée dans l'équipe d'Evans Lagudah (Moullet et al. 1999) au CSIRO ('Commonwealth Scientific and Industrial research Organization Plant Industry') en Australie. Son criblage a permis d'isoler un clone contenant le locus Ha. Ce clone, fourni par Sadequr Rahaman avec qui nous collaborions, a été séquencé sur des fonds Génoplante également.

I.2.3 Résultats comparatifs

Les premiers résultats sur le locus concernent la zone précise du locus Ha, définie comme celle contenant les gènes *Pina*, *Pinb* et *Gsp-1* sur les génomes qui possèdent les trois gènes. Ils ont permis de déterminer quelles sont les bases moléculaires des événements évolutifs à l'origine de la structure actuelle du locus Ha dans les blés diploïdes et polyploïdes [**P8**]. Pour les génomes ne possédant plus que le gène *Gsp-1*, le locus (surligné en bleu sur la **Fig. 3**) a été délimité en se basant sur les premiers gènes identifiés en 5' et en 3' du locus. Ces résultats peuvent être synthétisés de la manière suivante :

- La **perte des gènes *Pina* et *Pinb* est due à une (ou plusieurs) large(s) délétion(s) (environ 80 kb pour les génomes analysés) qui semble indépendante pour les génomes A et B.** La région génomique délétée n'a pas été remplacée par une séquence de taille équivalente puisqu'uniquement 3 à 5 kb séparent les bordures de la délétion. Les séquences identifiées dans cet intervalle sont composées essentiellement de résidus d'éléments transposables (différents entre les génomes A et B) suggérant leur potentielle implication dans la délétion par recombinaison illégitime, phénomène déjà décrit chez les céréales dont les génomes sont riches à très riches en éléments transposables.

- La comparaison entre les deux génomes D (celui du blé tendre et celui d'*Ae. tauschii*) a révélé de **nombreux remaniements** incluant des insertions d'éléments transposables, des inversions et des délétions. Une analyse fine des bordures des remaniements a révélé la présence de motifs répétés à des positions précises. Ces motifs, analysés en regards de ce qui a déjà été décrit dans la littérature, nous ont permis de faire l'hypothèses que des évènements de recombinaison illégitime ont pu avoir lieu dans cette région. **L'ensemble suggère que les génomes des blés évoluent rapidement,**

notamment par recombinaison illégitime, qui n'implique pas uniquement les séquences des éléments transposables.

Dans un deuxième temps, nous avons analysés la microcolinéarité au niveau des séquences étendues en 3' et 5' du locus Ha et montré que le niveau de **conservation de microcolinéarité est particulièrement réduit** à ce locus [**P13**]. Seuls deux groupes de gènes, qui sont retrouvés également dans le riz et l'orge, sont conservés. L'étude a également montré que les **gènes évoluent de façon très différentes** : certains sont bien conservés, d'autres ont subi des réarrangements (délétions ou duplications), d'autres suivent un processus de type 'birth and death' (multiplication du nombre de copies et dégénérescence d'une grande partie d'entre elles).

Ces travaux constituent l'une des **premières études de comparaison de séquences de grands fragments** d'un locus d'intérêt chez les céréales. Depuis, grâce à de nombreux efforts de recherches, étendus à l'ensemble du génome, la place prépondérante des séquences répétées (notamment des éléments transposables) ainsi que les impacts de la polyploïdie sur l'évolution des céréales (remaniements génomiques) ont été largement documentés, comme en atteste la vaste littérature sur ces sujets (voir ici quelques articles : Bolot et al. 2009; Feldman et al. 2012; Vicient and Casacuberta 2017; Parisod and Badaeva 2020).

Ces premiers travaux ont fait émerger un des questionnements qui m'anime encore aujourd'hui. En effet, le locus Ha présenté en figure 3 représente une infime fraction des génomes considérés et pourtant il met en évidence **des modes d'évolution des gènes très contrastés**. Coexistent à ce locus des gènes conservés et toujours aux mêmes positions et d'autres dont le nombre de copies varie entre génomes qui ont pourtant divergé depuis assez peu de temps. De plus, plusieurs copies de gènes incomplètes ou présentant des mutations non-sens ont été identifiées dans ce locus. Ces observations soulèvent des questions sur l'évolution des génomes et des gènes, qui vont devenir centrales dans la suite du document.

I.3. Autres activités, et réflexions générales

I.3.1. Génomique comparative au locus Ahd1 chez la canne à sucre

Comme évoqué plus haut, la canne à sucre est probablement l'une des espèces de *Poaceae* ayant le génome le plus complexe, car elle est hautement polyploïde et aneuploïde. Les cultivars modernes sont issus de l'hybridation interspécifique entre l'espèce polyploïde domestiquée *Saccharum officinarum* (x=10, 2n=8x=80) et l'espèce sauvage *S. spontaneum* (x=8, 2n=5x=40 à 16x=128). Ces deux espèces ont probablement chacune une origine autopolyploïde. Un haut niveau de colinéarité avait déjà été mis en évidence entre elles (Grivet et al. 1996) et avec le sorgho (Ming et al. 1998), espèce diploïde étudiée la plus proche de la canne à sucre.

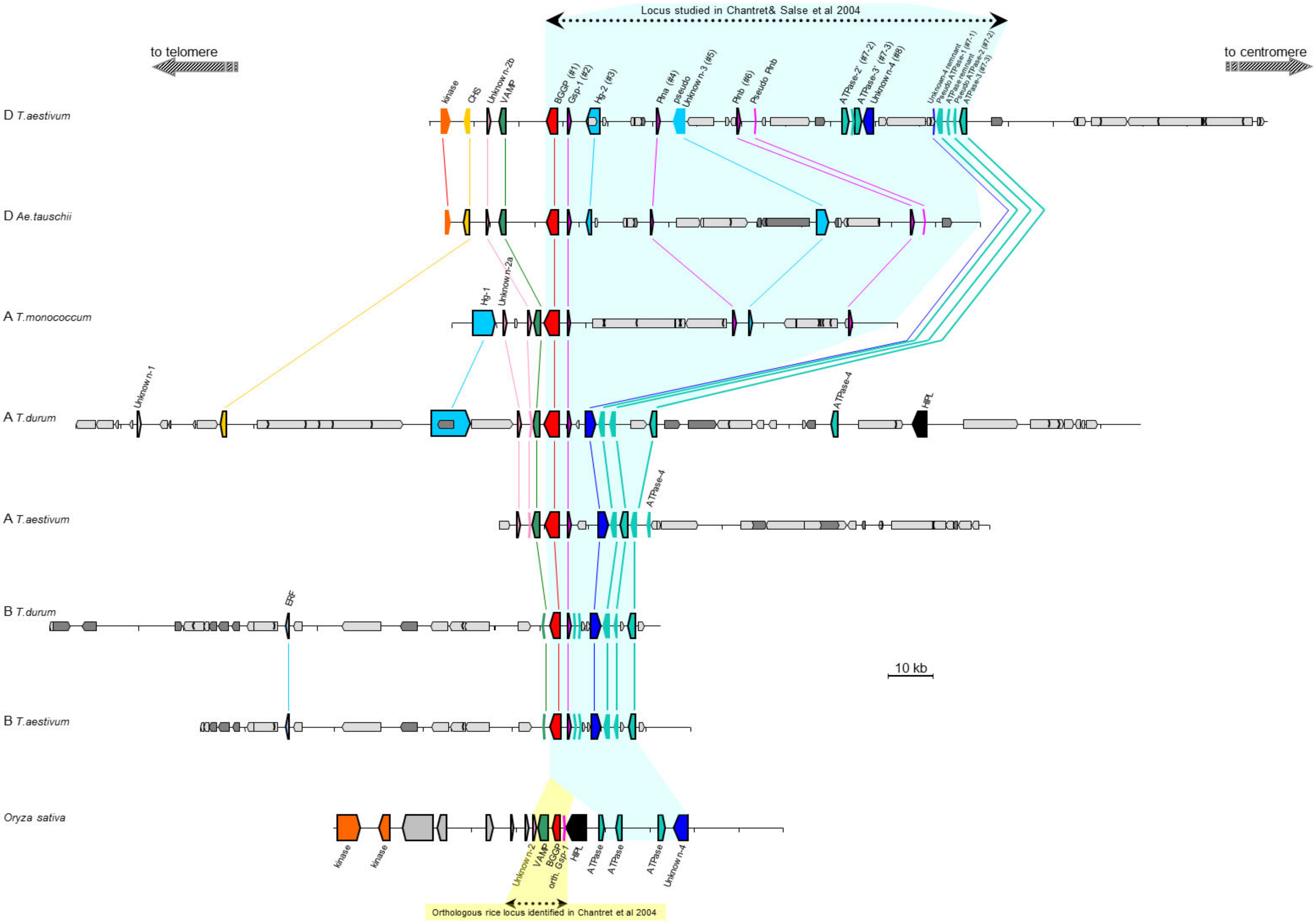

Locus studied in Chantret& Salse et al 2004
to telomere
to centromere
D T.aestivum
kinase
CHS
Unknow n-2b
VAMP
BGGP (#1)
Gsp-1 (#2)
Hg-2 (#3)
Pina (#4)
pseudo
Unknow n-3 (#5)
Pinb (#6)
Pseudo Pinb
ATPase-2' (#7-2)
ATPase-3' (#7-3)
Unknow n-4 (#8)
Unknown-4 remnant
Pseudo ATPase-1 (#7-1)
ATPase remnant
Pseudo ATPase-2 (#7-2)
ATPase-3 (#7-3)
D Ae.tauschii
A T.monococcum
Hg-1
Unknow n-2a
A T.durum
Unknow n-1
ATPase-4
HIPL
A T.aestivum
ATPase-4
B T.durum
ERF
B T.aestivum
10 kb
Oryza sativa
kinase
kinase
Unknow n-2
VAMP
BGGP
orth. Gsp-1
HIPL
ATPase
ATPase
ATPase
Unknow n-4
Orthologous rice locus identified in Chantret et al 2004

Figure 3 (ci-contre) Représentation schématique des 7 clones BACs de *Triticeae* ainsi que de 80 kb du chromosome 12 du riz (*Oryza sativa* ssp. *japonica*). La région étudiée dans l'article [**P8**] est surlignée en bleu. Les gènes potentiellement fonctionnels sont représentés par des larges flèches au contour noir. Les gènes tronqués ou les pseudogènes sont représentés par des flèches sans contour noir. Les éléments répétés de classe I et II sont représentés respectivement par des flèches en gris clair et en gris foncé. Figure issue de l'article [**P13**].

J'ai pu collaborer aux analyses du locus contenant le gène *Adh-1* (Alcohol deshydrogenase), un des premiers locus ciblés pour des analyses de génomique comparative entre le maïs, le sorgho et le riz (Tikhonov et al. 1999; Ilic et al. 2003). Ce projet avait pour objectif (i) de compléter avec la canne à sucre l'analyse de l'évolution de ce locus déjà bien documenté (ii) de décrire pour la première fois des séquences de grand fragment de cette espèce et (iii) de comparer la séquence de deux sous-génomes réunis dans une même espèce polyploïde. Deux BACs contenant le locus *Adh-1* provenant des deux génomes de la canne avaient été isolés, séquencés et analysés. L'un des résultats les plus marquants de ce travail a été que, contrairement à ceux obtenus dans la plupart des analyses de génomique comparative menées chez d'autres polyploides récents (blé, coton) ou plus anciens (maïs), **la conservation entre les deux génomes de canne à sucre s'est révélée extrêmement élevée**, aussi bien en termes de micro-colinearité (conservation parfaite) qu'au sein des espaces intergéniques qui se sont avérés très conservés (l'insertion de quelques éléments transposables mise à part) [**P11**].

I.3.2. Et depuis ?

Depuis cette période le niveau des connaissances a considérablement augmenté. Sans ambitionner de faire ici une synthèse de l'état actuel des connaissances, il est possible d'évoquer quelques éléments de réflexion issus de la littérature dans ces domaines.

Premièrement, le rythme de **publications de nouveaux génomes quasi complets** n'a cessé d'augmenter. Si plusieurs années furent nécessaires pour produire le premier génome d'Arabidopsis, il ne faut plus que quelques jours à l'heure actuelle, pour produire une séquence de qualité du même ordre de grandeur, grâce notamment à l'avènement des séquences de longs fragments (Michael et al. 2018). Néanmoins, les génomes polyploïdes ou hétérozygotes restent difficiles à assembler (Michael and VanBuren 2020). Aujourd'hui, le nombre de génomes complets de référence de plantes est estimé entre ~400 et 800 (Kitts et al. 2016; Thudi et al. 2021). Ainsi on sait que les tailles extrêmement variables des génomes ne sont pas corrélées à leur contenu en gènes (ploïdie mise à part) mais essentiellement à leur teneur en éléments transposables. En particulier chez les céréales à grands génomes comme le blé, les gènes ne sont pas répartis de façon homogène mais souvent en îlots et ils sont souvent contenus dans les extrémités télomériques des chromosomes. Au fur et à mesure, ces données ont permis de progresser dans la **reconstruction de l'histoire évolutive des génomes.** Ces recherches visent à modéliser les génomes ancestraux à différentes étapes clés de l'histoire des espèces et à construire des modèles de plus en plus intégratifs, qui retracent l'ensemble des évènements évolutifs (structuraux et moléculaires) qui ont pu conduire aux génomes tels qu'ils sont aujourd'hui (Salse 2016 ; Murat et al. 2017; Akoz and Nordborg 2019). En mettant en relation cette histoire avec l'évolution des traits d'histoire de vie, il est alors possible de comprendre comment se sont mises en place les grandes étapes du fonctionnement des organismes. La multiplication des sites et des bases de données hébergeant les séquences de génomes complets en est aussi une bonne illustration (NCBI, ensembl plant, plaBi database[11] …). L'ensemble de ces données a permis, par exemple, d'approfondir nos connaissances sur des mécanismes évolutifs d'importance majeure chez les plantes, comme la **polyploïdie** (quelques revues : Jiao et al. 2011; Soltis et al. 2015; Alix et al. 2017 ; Cheng et al. 2018; Mandakova and Lysak 2018), ou de modéliser les caractéristiques ancestrales des **premières fleurs** (Sauquet et al. 2017).

Que dire de la **portée de ces connaissances pour l'amélioration des plantes**, **dans le contexte agronomique et agroécologique actuel** ? Le premier lien vers l'application est celui de la notion de plante modèle, lié au concept de « biologie translationnelle » ou comment transposer les connaissances obtenues sur une espèce (modèle ou non) à une autre. La question récurrente de **l'orthologie** (évoquée en parallèle de celle de la **paralogie**), et de la **conservation de fonction** est au cœur de ces problématiques (Fitch 2000; Gabaldon and Koonin 2013) et fait l'objet de réflexions intenses (Linard et al. 2021). Le nombre de bases de données et d'outils dédiés aux prédictions d'orthologie est en augmentation constante, et doit relever le défi d'intégrer les données moléculaires produites à un rythme en croissance exponentielle (eggNOG (Huerta-Cepas et al. 2019), HieranoiDB, orthoDB (Zdobnov et al. 2021), greenphyl (Rouard et al. 2011) pour n'en citer que quelques-unes). Outre les questions relatives aux fonctions, c'est également sur les gènes 'orthologues' (*i.e.* gènes ayant divergé à partir d'un ancêtre commun uniquement par des évènements de spéciation) que reposent aussi les reconstructions phylogénétiques entre les espèces (Fitch 1970).

Le deuxième lien vers l'application est **l'extension de la génomique**, telle qu'elle était initialement abordée c'est-à-dire entre individus d'espèces différentes (interspécifique), **au compartiment intraspécifique.** Il est alors possible de mettre en parallèle deux champs de recherche qui abordent de façon différente la variabilité intraspécifique à l'échelle du génome : (1) l'identification d'allèles d'intérêt via la **génomique des populations** avec des méthodes comme la génétique d'association (GWAS) et (2) les analyses de type **pangénomique** (sur lesquelles nous reviendrons) qui permettent de caractériser la variabilité en 'contenu génomique' entre individus d'une même espèce. Une application des résultats provenant de ces deux champs de recherche réside dans la notion de **sélection assistée par la génomique** ou **GAS** ('Genomics-Assisted Breeding') passant par l'exploitation de la découverte allèlique (Varshney et al. 2021).

[11] https://plabipd.de/portal/fr/sequenced-plant-genomes

II. De la génomique structurale à l'évolution inter- et intraspécifique chez les Légumineuses : application aux gènes dupliqués

II.1. Changement d'unité : opportunités thématiques et méthodologiques

Les travaux de génomique structurale tels que je les avais abordés précédemment traitaient de l'évolution des génomes sous un angle 'descriptif'. Une question générale qui peut se poser sur ces résultats est de savoir si les changements observés ont été sélectionnés car procurant un avantage quelconque, ou s'ils ont été retenus (fixés) par hasard (dérive) au cours de l'évolution (Koonin 2009). En d'autres termes, **quelle est la part adaptative des modifications génomiques observées ?**

À plusieurs reprises au cours des années 2005 et 2006, j'ai pu rencontrer des membres de l'équipe 'structure de la diversité' de l'unité DIA-PC qui travaillait sur la **compréhension de mécanismes adaptatifs** (notamment liés à la **domestication**) **à l'échelle des populations dans plusieurs espèces ainsi qu'à l'échelle interspécifique, dans plusieurs complexes d'espèces**. Les corpus théoriques de la génétique des populations et de l'évolution moléculaire, étudiés dans cette équipe, offraient la possibilité de **poser la question du potentiel rôle adaptatif des modifications génomiques**. La donnée centrale pouvait rester la même : la séquence mais il ne s'agissait plus de comparer de grands fragments provenant d'espèces éloignées, obtenus sur un seul individu. C'est en comparant les séquences pour un même locus, soit sur plusieurs individus de la même espèce (échelle intraspécifique), soit sur des individus d'espèces différentes plus ou moins proches (échelle interspécifique), qu'il devient possible de poser la question du **rôle de la sélection dans le patron observé**. La démarche consiste notamment à élaborer des attendus sous des hypothèses précises et de comparer les patrons obtenus à ces attendus, ce qui est assez différent des approches exploratoires de la génomique structurale.

J'ai souhaité rejoindre cette équipe afin de développer **des projets articulant génomique et évolution**. À mon arrivée dans cette équipe, je me suis donc familiarisée avec ces nouvelles disciplines et j'ai mis en place de nouveaux questionnements, à l'interface entre ces deux champs disciplinaires. Cette immersion dans une équipe de biologie évolutive m'a permis de formuler différemment les questions qui avaient pu germer dans mon esprit au cours des années précédentes et surtout de concevoir une stratégie pour y répondre. Ces questionnements ont évolué et mûri et seront décrits dans la suite de cette synthèse jusqu'à mes activités actuelles.

En m'insérant dans cette équipe j'ai pu développer des recherches pour mieux comprendre, tout d'abord, **l'évolution des gènes dupliqués** (Partie II.3), puis la **dynamique évolutive des familles multigéniques** (Partie III) et plus récemment **l'épissage alternatif et son rôle potentiel dans l'adaptation** (Partie IV).

Enfin, un point important de ce changement est que, la dynamique d'équipe aidant, j'ai **consacré beaucoup plus de temps à l'encadrement d'étudiants**.

II.2. De la génomique structurale à la génomique des populations

II.2.1. Contexte

Lorsque j'ai rejoint l'équipe, l'avancée des connaissances dans le domaine de l'étude des génomes et de la variabilité nucléotidique permettait la mise en place des premiers projets de **reséquençage et d'analyse du polymorphisme.** Cette époque a marqué le **passage de l'étude de la diversité à l'aide de marqueurs neutres à celle de la diversité sur des gènes**, permettant ainsi de poser des questions sur **l'adaptation**. Dans ce contexte, l'équipe, et en particulier le groupe travaillant sur *Medicago truncatula* et ses espèces apparentées (**Encadré 1**) que je rejoignais, construisait des programmes de recherche en **génomique des populations** pour étudier la dynamique évolutive des populations et pour rechercher des **traces de sélection.** Ces approches reposent sur la **variabilité des séquences**, observable au sein d'un alignement, au niveau intra-spécifique, *i.e.* **analyse du polymorphisme**, mais également au niveau interspécifique, *i.e.* **analyse de la divergence**.

Trois projets ont été pour moi une formidable opportunité pour aborder ces nouvelles thématiques et y contribuer en apportant mes compétences sur le génome. Le projet **EAGLE**[12] dans lequel il était prévu de reséquencer finement plus de 50 fragments génomiques (d'environ 1kb) au sein d'une core collection d'une soixantaine d'individus représentant la diversité de l'espèce *M. truncatula* et d'une quinzaine d'espèces proches. Des gènes anonymes et des gènes candidats impliqués dans différentes fonctions d'intérêt chez les Légumineuses (comme par exemple les interactions entre plantes et rhizobium ou la phénologie de la floraison) avaient été choisis. L'approche de génomique des populations choisie était basée sur l'analyse contrastée des fragments anonymes et candidats. Cette utilisation des concepts de la génétique des populations était originale et représentait une étude pilote chez *Medicago*, complémentaire des approches de génétique d'association. Le projet ANR **Immunit-Ae**[13] dont l'un des volets était de rechercher, par génétique d'association, des gènes de résistance de *M. truncatula* au pathogène *A. euteiches*. Enfin, la **partie SP1 du projet ARCAD** (**Encadré 2**) et en particulier l'analyse des effets de la domestication sur les génomes des plantes et plus particulièrement sur la luzerne cultivée.

II.2.2. Contribution à la mise en œuvre des études de génomique des populations

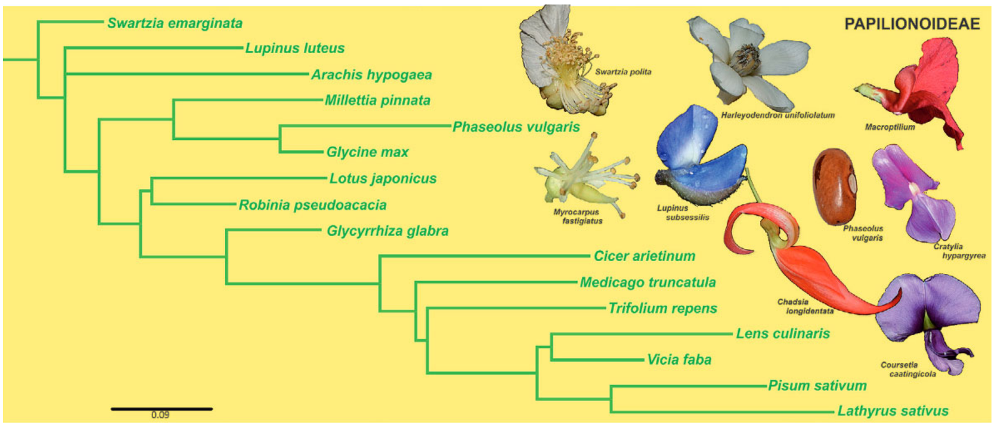
Mes compétences en génomique m'ont permis de m'impliquer dans ces projets dès les premières étapes **d'obtention des données de polymorphisme**. Il y a 15 ans, quand les nouvelles technologies de séquençage ne s'étaient pas encore développées, cela consistait à définir des zones cibles (notamment des gènes), des amorces pour les amplifier et séquencer le produit d'amplification (technologie de type Sanger), puis mettre en forme les séquences. Je me suis ainsi penchée sur la génomique de *M. truncatula* avec plaisir sachant que le séquençage de ce 'petit' génome (430 Mb) était en bonne voie. J'ai ainsi contribué à la recherche, à l'obtention et à **l'analyse des séquences** chez *M. truncatula*, puis à l'expertise et la mise en forme des séquences produites et aux analyses (alignement et descripteurs du polymorphisme). Lors de ces étapes, j'ai pu encadrer Karine Loridon (IE dans l'équipe), de la définition d'amorces jusqu'à la mise en forme des données, ainsi que dans la mise au point d'un set de SNP [**P16**].

[12] Pour 'Ecological and Association Genomics in LEgumes'

[13] Pour 'Diversité des composants génétiques et des mécanismes de résistance à *Aphanomyces euteiches* chez les légumineuses'

Encadré 1 *Medicago truncatula* et les légumineuses

La famille des légumineuses est l'une des plus importante en taille au sein des angiospermes. Elle contient autour de 19500 espèces, selon la classification de l' 'Angiosperms Phylogeny Group' (Azani et al. 2017), et inclut des espèces importantes pour la consommation humaine et animale, notamment de par leurs fortes teneurs en protéines dans leurs graines (soja, pois, haricot, pois chiche), ainsi que sous forme fourragère (luzerne ou trèfles). Elles ont la capacité de fixer l'azote atmosphérique via une interaction symbiotique avec des bactéries de type ryzobium. Cette propriété les rend également très précieuses d'un point de vue agroécologique comme alternative à l'apport d'engrais pour l'enrichissement en azote des sols.



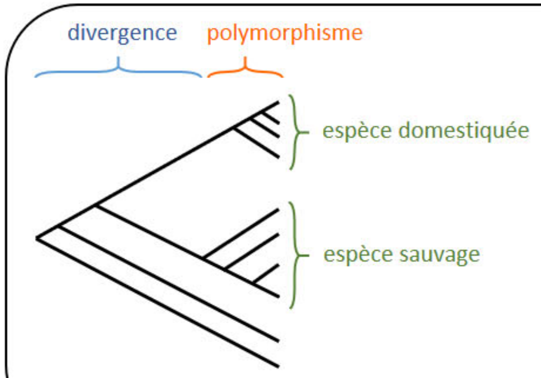
Phylogénie des papilionideae. Extrait de (Azani et al. 2017)

`Nom` : *Medicago truncatula*
`Genre` : Medicago
`Famille` : Fabaceae sens large (syn. Leguminosae = Legumineuses)
`Order` : Fabales
`Génome` : Diploïde, 430 Mb
`Cycle` : annuel
`Régime de reproduction` : majoritairement autogamme
`Aire de répartition` : pourtour méditerranéen
`Polymorphisme` : élevé

À la fin des années 2000, le choix de l'espèce *Medicago truncatula* comme plante modèle des légumineuses a émergé, pour l'étude moléculaire de la symbiose notamment mais également pour la génomique des légumineuses (Young et al. 2011; Pecrix et al. 2018).

Au début des années 2000, le groupe 'Madicago' de l'équipe s'est investi sur des questions relatives au fonctionnement et à la démographie des populations au sein de l'espèce *Medicago truncatula*. De gros efforts de prospections dans les populations naturelles avaient été faits depuis 1985 et l'étude de la diversité génétique à l'échelle de l'espèce a été publiée sur la base de génotypage par marqueurs microsatellites (Ronfort et al. 2006) et un centre de ressources génétique (CRB) a été créé. Ces résultats montraient notamment l'organisation en deux groupes génétiques (est et ouest) et proposaient la construction de core-collections emboitées. Le laboratoire avait également répertorié et acquis une grande expertise sur d'autres espèces appartenant au genre *Medicago*. Dans le cadre de l'étude du processus de domestication, des travaux sur la luzerne cultivée, espèce autotétraploïde, ont aussi été menés (Muller et al. 2006) et des prospections des espèces apparentées ont été réalisées.

Encadré 2 Le sous-projet 1 d'ARCAD « Comparative population genomics in wild and crop plants: a genome and phylogenetic wide approach »

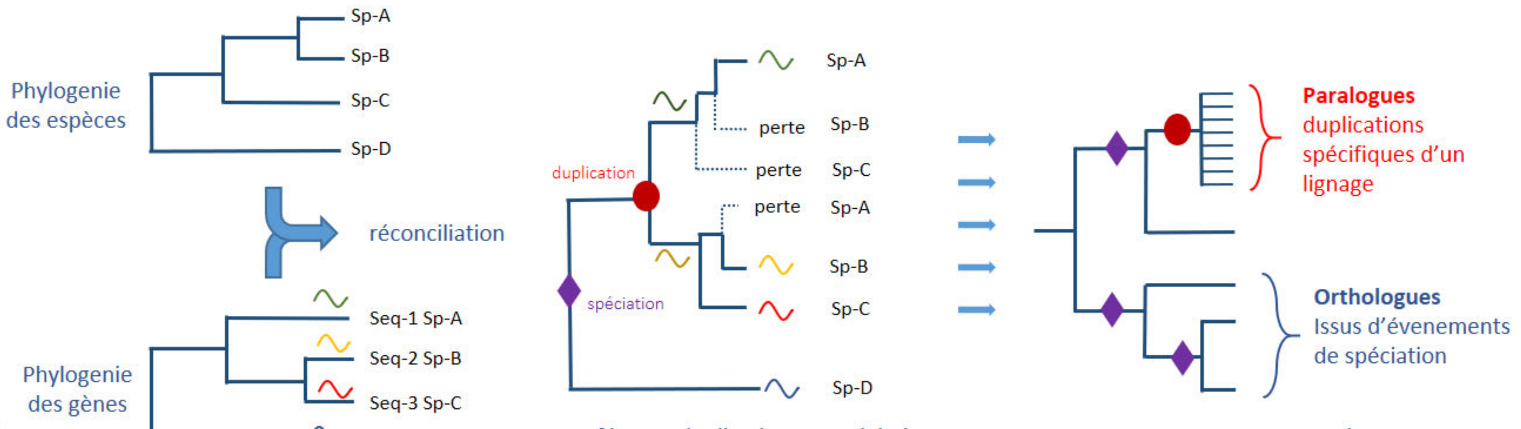


x 10 complexes d'espèces

WP3 : Génomique comparative du processus de domestication
Séquençage de 10 individus domestiqués, de 10 individus sauvages et de 2 outgroupes

Perte de diversité ?
Scénario démographique ?
Gènes de domestication ?

WP5 : Génomique comparative fonctionnelle
Identification des groupes de paralogues et d'orthologues au sein de la phylogénie de 10 angiospermes séquencés
Traces de sélection dans les paralogues vs. orthologues ?



Inférence duplications et spéciation

Identification paralogues/orthologues

II.2.2. Apports de ces nouvelles approches : tremplins pour l'émergence de nouveaux questionnements.

L'analyse du polymorphisme de séquence chez *M. truncatula* a mis en évidence un excès de mutations en faible fréquence, ce qui peut être interprété comme une signature **d'expansion démographique dans l'histoire de l'espèce** (De Mita et al. 2007). Une approche de type ABC (Approximate Bayesian Computation) a permis d'estimer des paramètres démographiques décrivant cette histoire [**P14**]. Pour identifier des traces de sélection dans les gènes et les fragments génomiques étudiés, nous avons construit, par simulations, des jeux de données reflétant l'histoire démographique de l'espèce. Le polymorphisme des gènes candidats a ensuite été comparé à ces attendus pour identifier des patrons de polymorphisme atypiques par rapport à l'histoire démographique neutre inférée. Ainsi, **deux gènes impliqués dans la nodulation et 4 gènes de la voie signalétique de la floraison ont montré des signatures de sélection positive** [**P14**]. Ce type de démarche a également été utilisé pour tester les gènes identifiés dans le cadre de la résistance à *Aphanomyces euteiches*.

Dans le cas de la **domestication** de la luzerne cultivée, cette même démarche a été appliquée pour, d'une part décrire les effets démographiques liés à la domestication, et, d'autre part, pour rechercher des loci pouvant révéler un écart aux attendus démographiques. Ces analyses ont donné lieu au stage de M2 de Maxime De Sario en 2014 que j'ai co-encadré avec Marie-Hélène Muller (CR dans l'équipe). Ce travail a été réalisé sur 7000 loci. L'ensemble des étapes de la méthode ABC et la validation des résultats ont été réalisés à l'aide de script 'à façon' utilisant la librairie egglib (De Mita and Siol 2012), occasion d'une collaboration avec Stéphane De Mita, un de ses auteurs. Les analyses ont permis de mettre en évidence une très faible perte de diversité dans le compartiment cultivé par rapport au compartiment sauvage chez la luzerne, apparaissant même compatible avec une **absence de goulot démographique** lors du processus de domestication. Ce résultat révèle une évolution, liée à la domestication, très atypique par rapport à ce qui est décrit pour la vaste majorité des espèces végétales domestiquées (Glemin and Bataillon 2009; Meyer and Purugganan 2013).

Cette période m'a permis de publier plusieurs résultats et ressources [**P14**, **P16**, **P17**, **P19**], d'encadrer un étudiant en master et de développer de nombreuses collaborations, notamment avec la communauté travaillant sur *Medicago truncatula*, au sein de l'équipe (Joëlle Ronfort, Marie-Hélène Muller, Stéphane De Mita, Concetta Burgarella, Jacques David, Jean-Marie Prosperi), en France (Sylvain Glémin) et à l'étranger (cf. liste des collaborations p.10 pour une liste exhaustive). Grâce à ces collaborations étroites au sein de l'équipe et à mon implication dans ces projets, j'ai pu élargir mes connaissances et appréhender les méthodes d'analyse de la **diversité génétique**, en particulier en prenant en compte l'information contenue dans le polymorphisme de séquence. J'ai pu, en particulier, comprendre comment se construisent les **tests de sélections en génomique des populations**. C'est à cette période également que je me suis initiée aux méthodes de **recherche de traces de sélection dans les génomes**. Grâce à ces travaux, j'ai pu commencer à **formuler des questions articulant la génomique structurale et la génomique des populations et surtout à entrevoir quels types de données et de méthodes je pouvais mobiliser pour y répondre**. La première question à laquelle je me suis attelée est celle du rôle des duplications de gènes.

II.3 Evolution de gènes dupliqués dans le genre *Medicago*

La progression des connaissances en génomique citées dans la Partie I a mis en lumière le **rôle crucial des évènements de duplications dans les génomes des plantes**. Dès l'étude du locus Ha (Partie I) la coexistence, dans une même région génomique, de gènes montrant des trajectoires évolutives si différentes m'avait intriguée. Parmi les évènements de duplication on peut mentionner ceux impliquant la duplication de génomes entiers, c'est-à-dire la polyploïdisation, qui ont eu un rôle majeur dans l'histoire des angiospermes (Adams and Wendel 2005). D'autres types de duplications sont aussi très fréquents, notamment chez les plantes : on parle de duplications segmentales lorsque des portions entières de chromosome se dupliquent ou de duplications en tandem lorsque de plus petites portions se dupliquent et se retrouvent côte à côte dans le génome. Une des conséquences de ces évènements est qu'à un temps t dans un génome donné, plusieurs copies du même gène peuvent coexister. Se posent alors de multiples questions : quelle est la **dynamique de fixation de ces duplications** ? **quel rôle a la sélection dans ce processus de fixation** ? **comment ces copies de gènes, initialement identiques, évoluent-elles** ? **quelles sont les conséquences à long terme de ces duplications** ? **quel rôle jouent-elles dans l'adaptation des plantes à leur environnement** ?

Dans la littérature de nombreux modèles ont été proposés pour décrire les différents destins possibles d'une paire de gènes paralogues, faisant ou non intervenir la sélection naturelle, et considérant ou non la phase pendant laquelle la duplication ségrège encore dans l'espèce. Une première formalisation de l'évolution des gènes dupliqués a été faite par Ohno il y a plus de 50 ans (Ohno 1970). Aujourd'hui, une vaste littérature traite du devenir des gènes dupliqués, décrivant aussi bien des modèles théoriques d'évolution (au niveau populationnel ou à de grandes échelles de temps (Lynch and Conery 2000; Walsh 2003; Innan 2009; Innan and Kondrashov 2010)) que des analyses de données empiriques (pour ne citer que quelques études incluant plusieurs espèces : Jiang et al. 2013; Panchy et al. 2016; Defoort et al. 2019; Qiao et al. 2019). Ainsi, plusieurs modèles sont devenus très populaires, comme le modèle « birth and death » qui est particulièrement adapté pour les gènes impliqués dans les interactions de type hôte/parasite et qui décrit un rythme rapide alternant duplication et perte de fonction des copies (Michelmore and Meyers 1998) (cf. Partie III), ou le modèle de « sub-fonctionnalisation » qui prédit qu'une paire de gènes dupliqués 'se répartit' les fonctions du gène ancestral par accumulation différentielle de mutations faiblement délétères et deviennent tous deux indispensables au bon fonctionnement de la plante (Innan and Kondrashov 2010).

Une question qui reste néanmoins très débattue dans ce contexte est de savoir **quel est le rôle de la sélection dans ces différents scénarii et dans quelle mesure les duplications de gènes ont un rôle dans l'adaptation** ? En effet, les preuves formelles du rôle adaptatif des duplications à l'échelle des génomes ne sont pas si nombreuses. Grâce aux approches méthodologiques décrites ci-dessus, en particulier la recherche de traces de sélection, j'ai entrepris d'aborder la question de **l'évolution des gènes dupliqués** en utilisant l'information portée dans leurs séquences, aux niveaux intra- et inter-spécifiques, pour amener des éléments de réponse à cette question.

Dans le cadre du stage de M2 de Joan Ho-Huu que j'ai co-encadré avec Joëlle Ronfort, nous avons réalisé l'analyse de l'évolution moléculaire de trois paires récentes de gènes dupliqués (deux paires de polygalacturonases *Pg11-Pg3* et *Pg11a-Pg11c* et une paire de transporteurs d'auxine *Lax2-Lax4*), chez *M. truncatula* et apparentées. Ces gènes ont été choisis car ils sont potentiellement impliqués dans le processus symbiotique. Les séquences de ces gènes ont été obtenues sur 17 espèces appartenant au genre *Medicago*. Nous avons recherché des traces de sélection en nous basant sur les modèles d'évolution des codons, et les rapports entre les taux de substitutions synonymes et non synonymes inférés par maximum de vraisemblance (Goldman and Yang 1994; Yang 1998). Nous avons cherché trois types différents de signatures de sélection : (i) la divergence du niveau de contraintes entre les gènes paralogues (ii) l'occurrence de sélection positive et (iii) le relâchement transitoire de contrainte après la duplication. Pour les trois couples, les pressions de sélection se sont avérées différentes entre les deux paralogues. Nous avons également trouvé **des sites sous sélection positive** chez *Pg11* alors que *Pg3* est principalement sous **sélection purificatrice**. Les deux gènes de la paire de paralogues la plus récente *Pg11a-Pg11c* montrent tous les deux des sites sous **sélection positive mais d'intensité différente**. Les gènes *Lax2* et *Lax4* sont tous les deux sous sélection purificatrice mais nous avons détecté un **relâchement transitoire de sélection après la duplication** [**P15**].

Utiliser la **recherche de signatures moléculaires de sélection dans des gènes dupliqués pour déterminer quel est leur destin évolutif** était une démarche assez originale chez les plantes. Chez les animaux, notamment chez les mammifères, il existe davantage d'études consacrées à l'analyse les pressions sélectives auxquelles ont été soumis les gènes dupliqués. Ces analyses sont souvent faites sur des gènes dupliqués par le mécanisme de rétrotransposition (un ARNm est rétrotranscrit et s'insère dans le génome) qui est prépondérant dans ces organismes (revue dans Casola and Betran 2017) et facilite l'identification des duplications, car la copie néoformée n'a plus d'introns. Ces études montrent aussi que les gènes (rétro)dupliqués peuvent être soumis aussi bien à des épisodes de sub- que de néofonctionnalisation à différents stades de leur évolution. Chez les plantes, des travaux plus récents ont été publiés dans lesquelles la divergence entre les gènes dupliqués est décrite, et parfois modélisée également, mais majoritairement à partir des profils d'expression ou des modifications épigénétiques (Roulin et al. 2013; Wang et al. 2016). Finalement encore assez peu d'études évaluent précisément le régime sélectif auquel sont soumis les gènes dupliqués en se basant sur les signatures moléculaires telles que les contrastes entre les sites non-synonymes et synonymes.

Ces premiers travaux ont constitué une première application des **approches de génomique des populations et d'analyse de la divergence pour répondre à des questions liées à la structure des génomes et à leur évolution**. Ils ont montré l'intérêt de ces méthodes et leur puissance pour documenter des trajectoires évolutives complexes. Nous avons ainsi pu mettre en évidence des **traces de sélection positives dans des gènes dupliqués récemment** attestant, à petite échelle, du rôle des duplications dans l'adaptation.

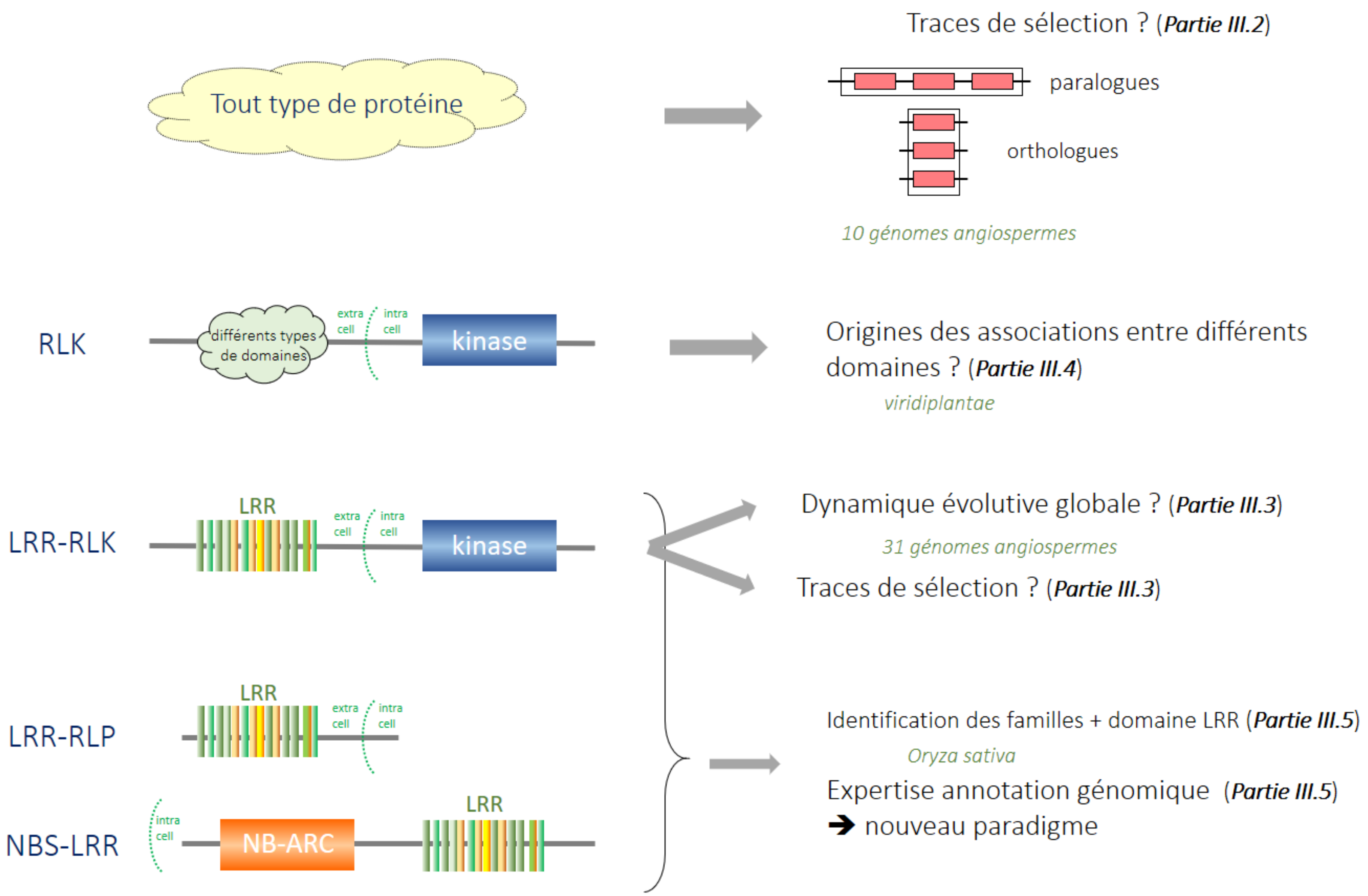

Tout type de protéine
Traces de sélection ? (*Partie III.2*)
paralogues
orthologues
10 génomes angiospermes
RLK
différents types de domaines
extra cell
intra cell
kinase
Origines des associations entre différents domaines ? (*Partie III.4*)
viridiplantae
LRR-RLK
LRR
extra cell
intra cell
kinase
Dynamique évolutive globale ? (*Partie III.3*)
31 génomes angiospermes
Traces de sélection ? (*Partie III.3*)
LRR-RLP
LRR
extra cell
intra cell
NBS-LRR
intra cell
NB-ARC
LRR
Identification des familles + domaine LRR (*Partie III.5*)
Oryza sativa
Expertise annotation génomique (*Partie III.5*)
➔ nouveau paradigme

Figure 4 Représentation schématique des différentes familles de gènes étudiées, incluant les espèces utilisées, la principale question abordée et la partie du rapport dans laquelle les résultats sont détaillés.

III. Evolution de familles de gènes

III.1. Contexte

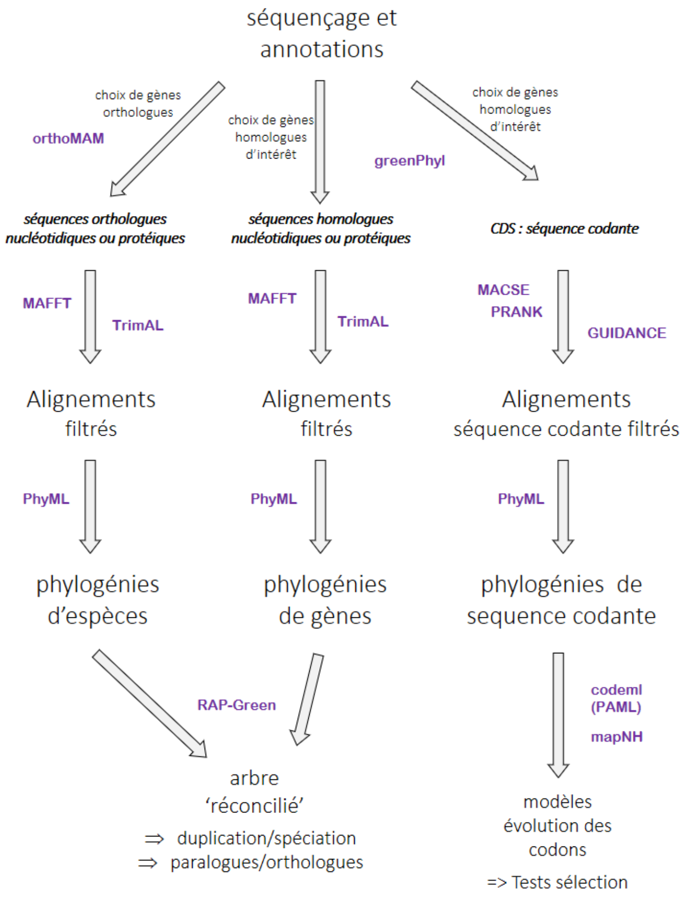
Les génomes des organismes supérieurs contiennent tous des familles multigéniques. D'un point de vue évolutif, une famille multigénique peut être définie comme un ensemble de gènes qui dérivent, par duplication, ou spéciation si plusieurs espèces sont considérées, d'un même gène ancestral commun, autrement dit, des gènes qui présentent de l'homologie de séquence (protéique/nucléique) (Demuth and Hahn 2009). Quelle que soit la méthode utilisée pour identifier ces familles, la répartition des gènes d'une espèce donnée entre gènes uniques et gènes appartenant à une famille multigénique, ainsi que les tailles de ces familles, sont extrêmement variables entre les différents phylums ou grands clades de l'arbre de la vie. Chez les plantes et en particulier chez les plantes à fleurs (angiospermes), cette organisation en familles de gènes est très prononcée (Kejnovsky et al. 2009). Que ce soit par la polyploïdie, par les duplications de grands fragments chromosomiques ou de segments plus petits, ou la rétrotransposition, les génomes des angiospermes sont hautement dynamiques comparés à la majorité des autres groupes de plantes terrestres (Leitch and Leitch 2012). Dans ce contexte, l'étude de l'évolution de ces familles multigéniques est un champ de recherche actif (Nei and Rooney 2005; Shakhnovich and Koonin 2006).

Jusqu'ici et comme décrit dans la partie précédente j'avais abordé la question du rôle des duplications sur quelques gènes seulement, et au sein d'un petit complexe d'espèces. J'ai développé ces travaux dans deux directions (1) en considérant **l'ensemble des copies d'une famille multigénique** qui sont le produit des duplications successives subies par un gène donné, à l'échelle d'un génome, et (2) en **augmentant l'échelle de temps des analyses grâce aux données génomiques disponibles au sein des angiospermes**, voir au-delà. Considérer l'évolution des familles de gènes dans leur globalité est l'aboutissement logique des questionnements sur les gènes dupliqués. En effet les familles de gènes sont le produit de duplications successives, qui s'échelonnent au cours du temps selon des rythmes et des intensités très variables. Appréhender une famille de gènes dans son ensemble est néanmoins assez différent, d'un point de vue méthodologique notamment, que de considérer un ensemble de paires de gènes.

La disponibilité toujours croissante des données génomiques et le développement d'outils méthodologiques permettent, actuellement avec plus de puissance et à des échelles plus larges, (1) de **documenter les évènements évolutifs à l'origine de l'organisation observable aujourd'hui de familles de gènes** : depuis quand existe une famille donnée ? quand ont eu lieu les expansions du nombre de ses copies ? et dans quelles branches phylogénétiques ? (2) d'**aborder la question du rôle adaptatif de ces duplications** en recherchant des traces de sélection notamment.

Pour aider la lecture des parties suivantes, j'ai représenté de façon schématique sur la **figure 4** les différents types de familles que j'ai étudiées, sur quel échantillonnage d'espèces, et les questions qui ont été adressées. Dans l'**encadré 3** sont mentionnées les différentes méthodes utilisées dans ces analyses et comment elles peuvent s'articuler.

Encadré 3 Schéma représentant l'articulation entre les différentes méthodes utilisées (Partie III).


séquençage et annotations
choix de gènes orthologues
orthoMAM
choix de gènes homologues d'intérêt
choix de gènes homologues d'intérêt
greenPhyl
séquences orthologues nucléotidiques ou protéiques
séquences homologues nucléotidiques ou protéiques
CDS : séquence codante
MAFFT
TrimAL
MAFFT
TrimAL
MACSE
PRANK
GUIDANCE
Alignements filtrés
Alignements filtrés
Alignements séquence codante filtrés
PhyML
PhyML
PhyML
phylogénies d'espèces
phylogénies de gènes
phylogénies de sequence codante
RAP-Green
codeml (PAML)
mapNH
arbre 'réconcilié'
⇒ duplication/spéciation
⇒ paralogues/orthologues
modèles évolution des codons
=> Tests sélection

Annotations : étapes permettant d'identifier les éléments d'une séquence nucléotidique (exons, introns, parties codantes) ou les domaines protéiques fonctionnels dans une séquence protéique, en utilisant des approches de prédiction *de novo* ou par comparaison de séquences.

Comparaisons de séquence et alignement. Lorsqu'une paire de séquences (nucléiques ou protéiques) est considérée deux approches existent (1) les méthodes d'alignement 'locales' qui permettent d'identifier des régions homologues entre deux séquences comme BLAST qui, à partir d'une séquence requête, identifie des régions homologues dans des séquences cibles, au sein d'une base de données, (2) les méthodes d'alignement globales qui consistent à insérer des gaps de façon à ce qu'à chaque caractère de la première séquence corresponde un caractère de la seconde (ou un gap) pour révéler au mieux l'homologie potentielle existant entre deux séquences, en maximisant un score calculé à partir de coût définis sur les correspondances entre sites, les différences et les gaps. **Les alignements multiples**, qui sont une représentation de l'homologie entre les sites à l'échelle de plusieurs séquences, sont beaucoup plus compliqués à construire d'un point de vue algorithmique. De très nombreuses méthodes existent et les recherches sont encore actives dans ce domaine. Le '**nettoyage**' **d'un alignement** consiste à enlever ou masquer des parties d'un alignement parce qu'elles sont jugées de mauvaise qualité. Il existe également de nombreux outils qui réalisent ces nettoyages mais le bénéfice n'est pas toujours si évident. J'ai co-écrit un chapitre d'ouvrage sur les forces et les limites des alignements multiples et des méthodes de filtrage avec Vincent Ranwez [**P29**].

Phylogénie : À partir des alignements multiples, il est possible d'inférer une phylogénie. Plusieurs approches existent (1) les arbres de distances qui sont construits sur les matrices de distances obtenues pour chaque paire de séquence (2) les méthodes qui utilisent le maximum de vraisemblance pour estimer des paramètres sous-jacents à un modèle évolutif donné (3) les méthodes Bayesiennes. Ces deux dernières sont les plus utilisées et sont plus aptes à révéler l'histoire évolutive des séquences analysées.

Encadré 3 Méthodes utilisées *(suite)*

Réconciliation : Cette étape consiste à comparer un arbre phylogénétique de gènes obtenus dans x espèces avec l'arbre phylogénétique des espèces correspondantes. Ces méthodes permettent de déterminer combien et dans quelles branches de l'arbre se situent les évènements de duplication et de perte de gènes et ainsi d'identifier les groupes de gènes orthologues et les groupes de gènes paralogues, au sein d'un échantillon d'espèce.

Tests de sélection (contexte phylogénétique/interspécifiques) : Les méthodes les plus utilisées se basent sur les modèles d'évolution des codons en utilisant notamment le paramètre ω qui est le rapport entre les taux de substitutions non synonymes et synonymes (ω =dN/dS). Les paramètres, dont ce rapport ω, sont estimés le long des branches de l'arbre phylogénétique et/ou à chaque site de la séquence par maximum de vraisemblance (Yang 2007). Les alignements utilisés pour pouvoir appliquer ces méthodes doivent être nucléotidiques, comporter uniquement des **séquences codantes** et respecter la **phase de lecture** d'un gène, c'est-à-dire aligner des codons. Parmi les outils permettant de réaliser ce type d'alignement, on peut citer MACSE qui a la spécificité de prendre en compte de possibles décalages de phase et codons stop (Ranwez et al. 2011) [**P27**, **P30**].

III.2. Sélection positive dans des familles de gènes sans a priori fonctionnel

En inférant, par réconciliation phylogénétique, les duplications, les rétentions et les pertes des gènes au sein des génomes des plantes, une très forte diversité est observée entre les familles multigéniques en termes de nombre de copies, de vitesse de duplication et de taux de rétention observés. Cette diversité est la conséquence des forces évolutives en jeu lors de ces processus de duplication, de rétention et/ou de perte de gènes. Dans le cadre du SP1 du projet ARCAD, en particulier au sein du WP5 (**Encadré 2**), nous nous sommes posés la question du rôle de l'adaptation dans cette diversité. En effet, il existe des éléments indiquant un rôle adaptatif du maintien des gènes dupliqués, et en particulier lorsque se produit une **expansion rapide du nombre de copies dans un lignage spécifique** (Hanada et al. 2008). Ainsi nous avons cherché à déterminer si, et dans quelle mesure**, les duplications récurrentes spécifiques des lignages** (et suivies de la rétention des copies résultantes) **sont adaptatives**. En faisant l'hypothèse que des traces de sélection positive sont la preuve qu'un processus adaptatif a agi dans l'histoire d'une famille de gènes, nous avons cherché à déterminer si des traces de sélection positive étaient plus fréquemment observées dans les gènes dupliqués que dans les gènes non dupliqués, en considérant 10 génomes complets d'angiospermes (**Fig. 5**).

Nous avons utilisé la base de données GreenPhylDB (contenant les phylogénies des familles protéiques pour plusieurs génomes complets de plantes (Rouard et al. 2011)) pour extraire d'une part des groupes d''ultraparalogues' définis comme issus uniquement d'évènement de duplication et, d'autre part, des groupes de 'superorthologues' définis comme issus uniquement d'évènements de spéciation et ceci pour les dix génomes d'angiospermes les mieux annotés au moment de l'étude (**Fig. 5**). Les étapes suivantes ont consisté à aligner ces groupes d'ultraparalogues et de superorthologues en utilisant des méthodes permettant de conserver la phase de lecture (alignement de codons avec l'option 'codon' du logiciel PRANK (Loytynoja and Goldman 2005)), à nettoyer de façon automatique pour éliminer les parties mal alignées qui sont souvent à l'origine de faux positifs lors de la recherche de traces de sélection (outil GUIDANCE (Penn et al. 2010)) et à inférer les phylogénies par maximum de vraisemblance avec le logiciel PhyML (Guindon et al. 2010). Les recherches de traces de sélection ont été faites en utilisant les modèles d'évolution des codons via le rapport entre les taux de substitutions non-synonymes et synonymes ($dN/dS=\omega$), sur (i) les codons, en utilisant le logiciel codeml (programme PAML (Yang 2007)), lui-même intégré dans la bibliothèque egglib (De Mita and Siol 2012) et (ii) les branches, en utilisant le logiciel mapNH (Dutheil et al. 2012; Romiguier et al. 2012) (**Fig. 5**).

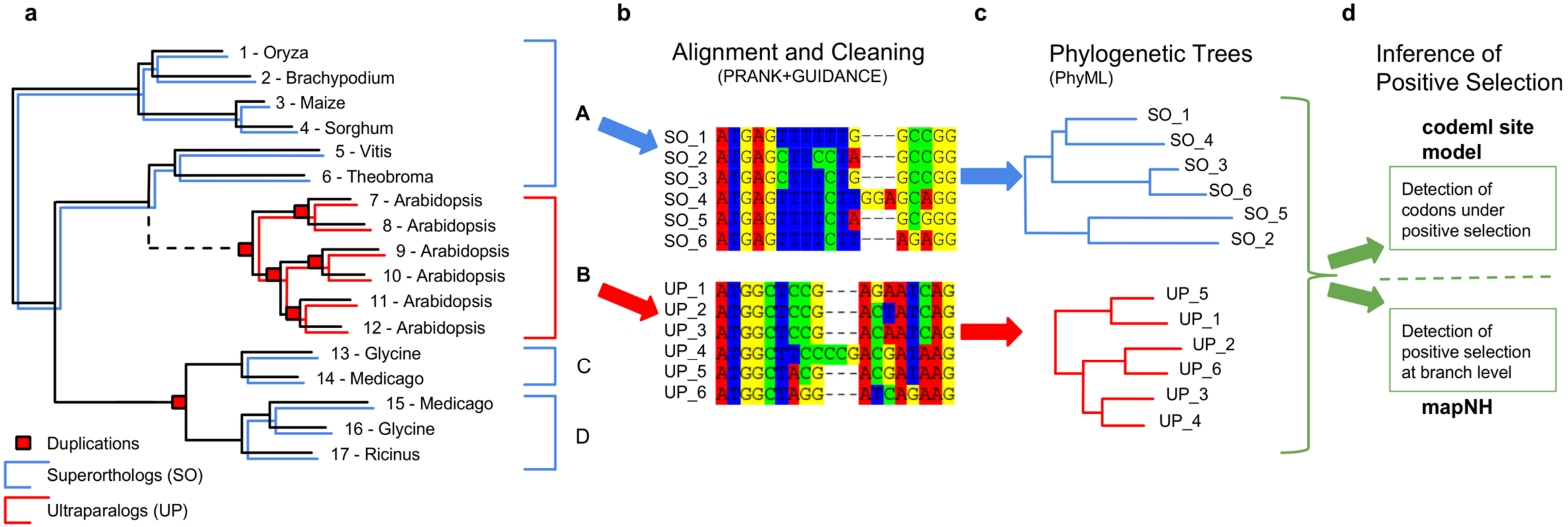

a
1 - Oryza
2 - Brachypodium
3 - Maize
4 - Sorghum
5 - Vitis
6 - Theobroma
7 - Arabidopsis
8 - Arabidopsis
9 - Arabidopsis
10 - Arabidopsis
11 - Arabidopsis
12 - Arabidopsis
13 - Glycine
14 - Medicago
15 - Medicago
16 - Glycine
17 - Ricinus
A
B
C
D
Duplications
Superorthologs (SO)
Ultraparalogs (UP)
b
Alignment and Cleaning
(PRANK+GUIDANCE)
SO_1 ATGAGTTTTTTG---GCCGG
SO_2 ATGAGCTTCCTA---GCCGG
SO_3 ATGAGCTTTCTG---GCCGG
SO_4 ATGAGTTTTCTTGGAGCAGG
SO_5 ATGAGTTTTCTA---GCGGG
SO_6 ATGAGTTTTCTT---AGAGG
UP_1 ATGGCTCCG---AGAATCAG
UP_2 ATGGCTCCG---ACTATCAG
UP_3 ATGGCTCCG---ACAATCAG
UP_4 ATGGCTTCCCCGACGATAAG
UP_5 ATGGCTACG---ACGATAAG
UP_6 ATGGCTAGG---ATCAGAAG
c
Phylogenetic Trees
(PhyML)
SO_1
SO_4
SO_3
SO_6
SO_5
SO_2
UP_5
UP_1
UP_2
UP_6
UP_3
UP_4
d
Inference of Positive Selection
codeml site model
Detection of codons under positive selection
Detection of positive selection at branch level
mapNH

Figure 5 Représentation de l'enchaînement des étapes utilisées pour rechercher des traces de sélections dans les (ultra)paralogues et les (super)orthologues au sein des Angiospermes. (a) Exemple d'un arbre phylogénétique protéique disponible dans GreenPhylDB (Rouard et al. 2011). Les séquences d'Arabidopsis 7 à 12 (cluster B) ne sont reliées que par des évènements de duplication (nœuds représentés avec des carrés rouges) et sont donc des 'ultraparalogs' (=UP; lignes rouges). Les séquences reliées uniquement par des évènements de spéciation sont des 'superorthologues' (=SO; lignes bleues). Ce sont les séquences 1 à 6 (cluster A), 13 et 14 (cluster C), 15 à 17 (cluster D). Par exemple, les séquences 13 et 15 sont paralogues (car elles sont reliées par des duplications) mais pas ultraparalogues (car un évènement de spéciation s'est produit après cette duplication). Nous n'avons utilisé que les clusters qui ne contenaient que 6 séquences au minimum (clusters A et B) pour les analyses. (b) Les séquences des CDS correspondants sont téléchargées pour chaque cluster et alignées. (c) Les arbres phylogénétiques ont été inférés pour chaque alignement (d) Les traces de sélection positive ont été inférées sur les codons et sur les branches dans tous les alignements.

Brièvement, nous avons trouvé **dans les ultraparalogues, près de 50 fois plus de gènes pour lesquels des sites sous sélection positive sont observés, que dans les superorthologues** (Tableau 1). Les sites sous sélection positive ont fait l'objet d'une vérification manuelle, afin d'éviter les faux positifs, fréquemment observés dans les alignements de mauvaise qualité. L'étude des vitesses d'évolution sur les branches des arbres phylogénétiques montre également en moyenne une valeur significativement plus élevée de ω pour les paralogues (ω =0.53) que pour les orthologues (ω =0.27) ainsi qu'un nombre de branches ayant un ω supérieur à 1.2 plus élevé pour les paralogues que pour les orthologues (Tableau 1).

Tableau 1 Nombre de clusters sous sélection positive. Clusters testés avec les modèles 'sites' (colonne 2, 3 et 4) et 'branches' (colonnes 5, 6 et 7). UP : 'ultraparalogues', SO : 'superorthologues'.

	Nombre de clusters (modèles sites)			Nombre de clusters testés (modèles branche)		
	testés	avec des codons sous sélection positive	avec des codons sous sélection positive après nettoyage manuel	testés	avec des branches dont le rapport ω>1	avec des branches dont le rapport ω>1.2
UP	1672	215 (12.9%)	90 (5.4%)	1589	819 (51.5%)	582 (36.6%)
SO	1356	4 (0.3%)	0 (0.0%)	1257	49 (3.9%)	24 (1.9%)

Ces résultats montrent clairement que, dans les génomes des angiospermes testés, la duplication de gènes peut effectivement être suivie d'épisode de sélection positive, induisant la fixation de mutations non synonymes à plus forte fréquence que les mutations synonymes. Ces mutations différencient ainsi les copies les unes des autres. **Les gènes dupliqués apparaissent comme un support sur lequel la sélection positive agit davantage que sur les gènes mono-copie**. L'identification d'une forte proportion de sites sous sélection positive dans les clusters d'ultraparlogues est cohérente avec les modèles de néofonctionnalisation qui prédisent une différenciation entre les copies dupliquées conduisant à l'émergence de nouvelle fonction. D'autre part, le relâchement de la sélection purificatrice, beaucoup plus important au sein des ulstraparalogues que des superorthologues, est, quant à lui, cohérent avec le modèle de subfonctionnalisation. Plusieurs forces évolutives modèlent le destin des gènes dupliqués, confirmant que ces derniers sont bien un **support important pour l'innovation fonctionnelle**. Ce travail a également mis en évidence l'importance de la qualité des données utilisées pour réaliser ces inférences, et en particulier la validation manuelle des sites sous sélection positive s'est avérée essentielle et a d'ailleurs été appréciée par les reviewers de l'article dans lequel nous avons publié ces résultats [**P18**].

Ce travail a été réalisé avec deux post-doctorants que je co-encadrais : Iris Fischer (post-doctorant ARCAD) et Jacques Dainat (post-doctorant Institut Agro) [**P20**], en collaboration avec Sylvain Glémin (Chercheur CNRS - ISEM), Jacques David (Prof. Institut Agro GE^{2}pop), Jean-François Dufayard (Chercheur CIRAD AGAP) et Vincent Ranwez (Prof. Institut Agro GE^{2}pop).

III.3. Sélection positive dans les récepteurs kinase de type LRR-RLK

Dans la partie précédente l'étude n'était pas centrée sur des gènes de fonction connue. Pour pouvoir interpréter des résultats d'évolution moléculaire en lien avec la fonction des gènes, j'ai orienté mes recherches sur une famille de gènes particulièrement intéressante chez les plantes car impliquée, entre autres, dans l'immunité : **les gènes codant pour des récepteurs kinase contenant des répétitions riches en leucine ou 'LRR' (LRR-RLK)** (**Fig. 4**)**.** Pour cela, j'ai développé une collaboration avec Anne Diévart (chercheur CIRAD, AGAP, équipe DAR), spécialiste de ces récepteurs à l'échelle moléculaire (Dievart and Clark 2003; Dievart and Clark 2004; Dievart et al. 2016). Les LRR-RLK contiennent trois domaines : un domaine LRR extracellulaire (N-terminal), un domaine transmembranaire et un domaine kinase (KD) (C-terminal) (**Fig. 4**). Sur la base des relations phylogénétiques entre les domaines kinases, 15 sous-groupes de LRR-RLK ont pu être définis chez *Arabidopsis* (Shiu et al. 2004; Lehti-Shiu et al. 2009). Ces gènes font l'objet de recherches intensives depuis le début des années 1990 [**P28**] car ils sont associés à des fonctions essentielles de signalisation cellulaire et de réponse aux stress environnementaux biotiques et abiotiques (Tang et al. 2010) – ils sont donc particulièrement pertinents à étudier dans le contexte actuel de transition climatique et agroécologique.

J'ai encadré Iris Fischer (bourse gouvernementale allemande) avec qui nous avons étudié la **dynamique évolutive de cette famille chez les angiospermes**. En se basant sur la phylogénie de 7554 gènes LRR-RLK provenant de 31 génomes entièrement séquencés de plantes à fleurs, nous avons pu documenter la dynamique complexe de l'évolution de cette famille. Nous avons montré que les différents sous-groupes de gènes ont des taux de duplication extrêmement contrastés. Les gènes ayant le plus fort taux de duplication se trouvent principalement dans les sous-groupes impliqués dans les interactions environnementales, notamment biotiques (perception et transduction du signal lors d'une agression par des organismes pathogènes). Ces analyses nous ont aussi permis d'identifier les branches de l'arbre phylogénétique, et les espèces, pour lesquelles un taux élevé de duplication pour certains sous-groupes, plus fort qu'en moyenne, pourrait être le marqueur d'un avantage sélectif (**Fig. 6**).

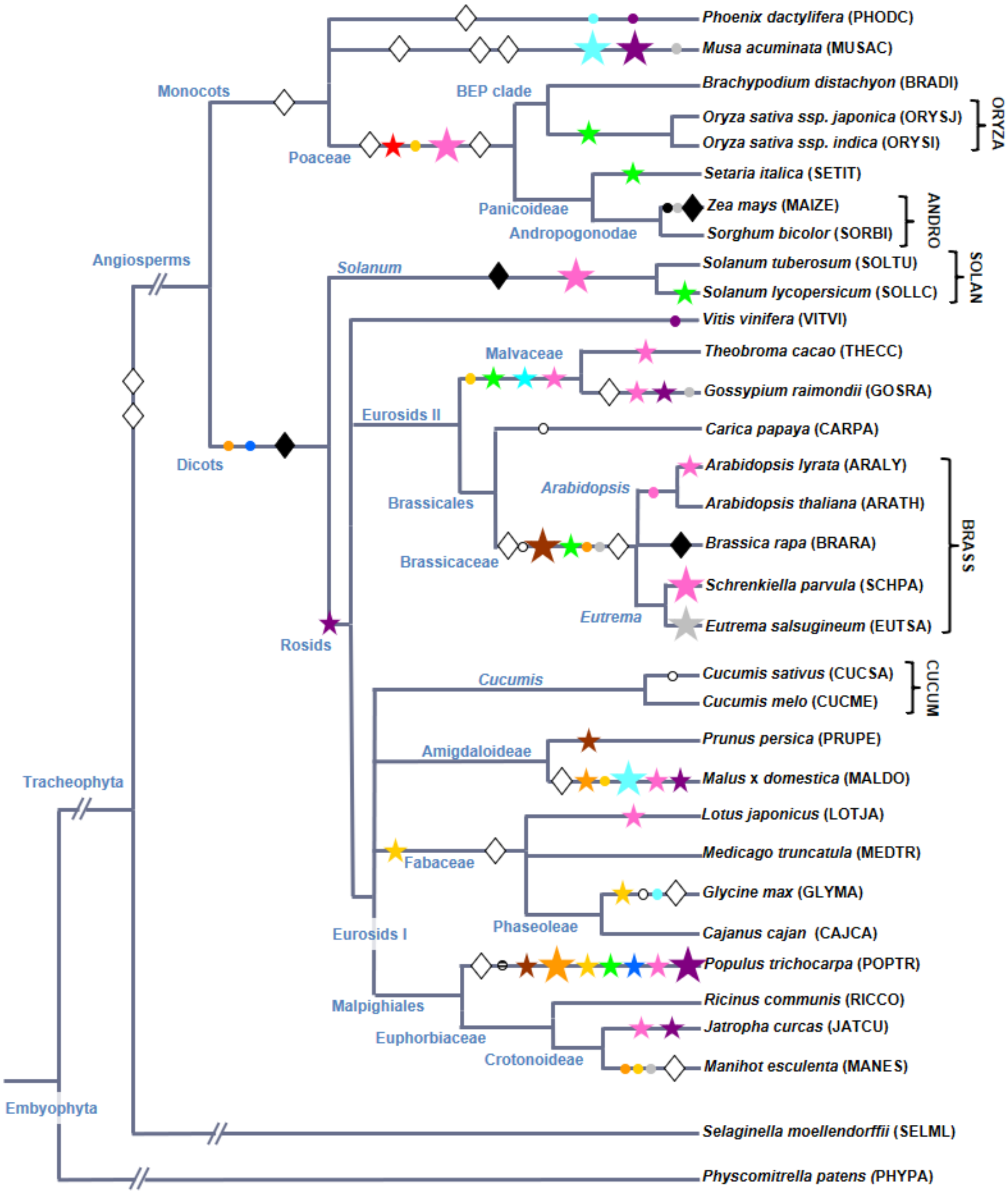


Figure 6 Arbre phylogénétique des 33 espèces étudiées. Les losanges blancs et noirs représentent les évènements de polyploïdisation décrits dans la littérature (en blanc doublement du génome, en noir triplement). Les points, les petites et les grandes étoiles, indiquent les évènements de multiplication respectivement de 2, entre 2 et 4, et de plus de 4 fois, du nombre de copies. Chaque couleur représente un sous-groupe différent de LRR-RLK [**P20**].

À partir de ces données, nous avons, dans un deuxième temps, étudié précisément le rôle de la sélection lors de l'expansion de cette famille. Nous avons extrait 75 groupes de LRR-RLK dupliqués récemment (paralogues identifiés spécifiquement au sein d'une seule et même espèce, représentant un total de 796 séquences) et 189 groupes de LRR-RLK orthologues (un gène par espèce, représentant un total de 1970 séquences). En utilisant des tests basés sur le ratio entre les taux de substitutions non-synonymes sur synonymes (dN/dS), nous avons recherché des traces de sélection dans ces alignements de gènes dupliqués récemment (paralogues) et de gènes orthologues (**Encadrés 2 et 3**). Nous avons pu détecter **des codons sous sélection positive dans 50 % des gènes appartenant à cette famille et dupliqués récemment.** Dans la Partie III.2, sur l'ensemble des autres familles, nous n'avions détecté des signatures de sélection positive qu'au sein de 12% des groupes de paralogues. Ce résultat met en **évidence le rôle prépondérant de la sélection positive dans l'évolution des duplications de LRR-RLK**. En outre, les domaines LRR, et plus particulièrement **quatre acides aminés du motif LRR**, se sont avérés être les principales cibles de la sélection positive (**Fig. 7a**). Ces acides aminés sont dans la première partie des motifs LRR. Dans la structure tridimensionnelle de ce domaine, ils se retrouvent dans la partie exposée au milieu externe (**Fig. 7b**). Enfin, nous avons observé que chez les Brassicaceae, plus de 10% des groupes d'orthologues sont absents, suggérant que d'importantes pertes de gènes ont eu lieu dans ces lignages, relativisant ainsi la position d'*Arabidopsis thaliana* comme seule espèce modèle pour des analyses fonctionnelles.

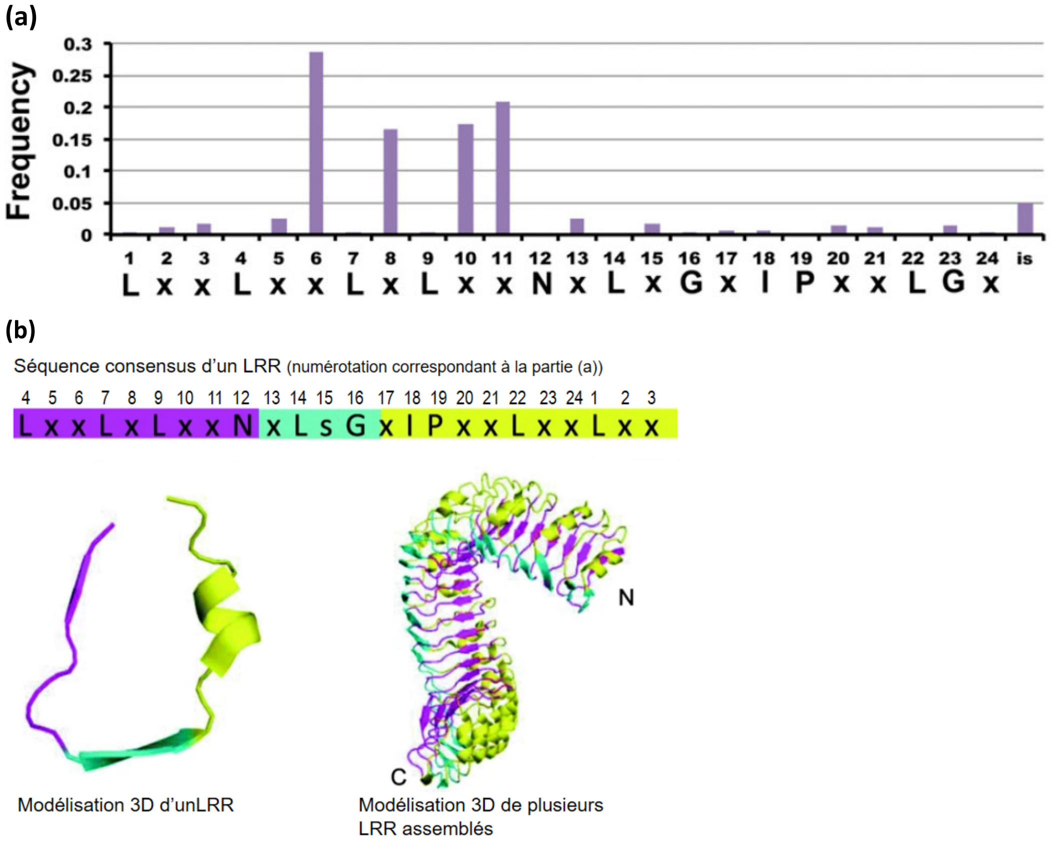


Figure 7 Structures protéiques d'un motif LRR et résultats des tests de sélection (a) fréquence à laquelle chaque acide aminé a été détecté sous sélection positive, dans un motif LRR, dont la séquence en acides aminés est représentée en abscisse ; L, Leucine ; N, Asparagine ; G, Glycine ; I, Isoleucine ; P, Proline ; x, n'importe quel résidu ; au sein du domaine LRR, les sites qui sont entre les motifs LRR sont notés 'is' pour 'ilots' (issu de [**P20**]). (b) représentation de la structure tertiaire d'un motif LRR à gauche, et de l'ensemble du domaine à droite. Le code couleur correspond à la séquence primaire du motif (en haut). La partie en violet est la partie interne de la super hélice. Issu de (Chen 2021).

Ces résultats [**P20**, **P23**] démontrent que **des évènements de sélection positive ont eu lieu plusieurs fois au cours de l'évolution de la famille multigénique des récepteurs kinase contenant un domaine LRR**. Ces évènements ont favorisé **la diversification du domaine LRR, en ciblant les parties protéiques impliquées dans les interactions moléculaires**, interactions qui ont lieu notamment avec les molécules émises par les pathogènes.

A la suite de ces résultats, avec ma collègue Anne Diévart, nous avons été sollicitées pour contribuer à l'analyse des données de séquençage du génome du chêne, coordonnée par Christophe Plomion (INRAE Bordeaux). Nous avons analysé à nouveau les gènes codant pour des récepteurs contenant des LRR, notamment ceux identifiés comme hautement dupliqués chez le chêne, et recherché des traces de sélection. **Sur 24 groupes de gènes identifiés issus de duplications récentes, 19 présentent des signaux significatifs de sélection positive, et 78% des sites sous sélection positive appartiennent aux domaines LRR**. Ces résultats, inclus dans la publication du génome du chêne [**P26**], montrent une fois de plus que cette famille, impliquée dans la perception de signaux moléculaires, évolue largement par duplication et diversification, cette dernière étant liée en grande partie à des évènements de sélection positive.

III.4. Analyse évolutive des récepteurs kinase chez les plantes

Les approches phylogénétiques que nous avons utilisées dans les parties précédentes pour documenter la dynamique évolutive des récepteurs kinase à LRR au sein des angiospermes peuvent être utilisées à une plus grande échelle de temps pour remonter à l' 'origine' de ces familles. Les récepteurs kinase contenant des LRR précédemment cités font partie d'une plus grande famille : celle des **récepteurs kinase** (RLK) (**Fig. 4**). Les RLK constituent l'une des plus grandes familles de gènes identifiée chez les plantes et elle s'est massivement développée au sein des plantes terrestres (Embryophyta ; (Lehti-Shiu et al. 2009; Lehti-Shiu and Shiu 2012)). Cette grande famille comporte des gènes qui possèdent, de façon schématique, un domaine kinase intracellulaire, un domaine transmembranaire et un domaine extracellulaire. Au-delà de la dynamique propre de chaque sous-famille, comme celle que nous avons étudiée pour les LRR-RLK, la grande diversité des combinaisons de domaines dans les récepteurs kinase, en particulier dans leur domaine extra-cellulaire, soulève plusieurs questions évolutives, qui concernent cette fois leur origine : **quand chacune de ces combinaisons est-elle apparue** ? Ces combinaisons sont-elles apparues **au même moment dans l'histoire évolutive ou à des périodes très distinctes** ? Chacune d'entre elles n'est-elle apparue qu'une seule fois, ou certaines sont-elles apparues plusieurs fois au cours de l'évolution (évolution convergente) ?

La disponibilité des données de séquences de génomes entiers, et cette fois incluant ceux d'espèces en dehors des angiospermes, nous a permis de rechercher systématiquement l'occurrence de ces combinaisons de domaines, en s'appuyant sur ce qui est connu de la phylogénie de ces lignages ancestraux. Nous avons ainsi pu montrer que la grande majorité de ces combinaisons de domaines sont apparues avant l'avènement des angiospermes et ont jalonné l'évolution des streptophytes (ou plantes vertes). Ces combinaisons ne sont pas apparues toutes en une seule étape mais graduellement et il est possible d'émettre des hypothèses quant aux relations entre l'apparition de certaines familles et la complexité morphologique, ou les mécanismes de développement et de signalisation/défense nécessaires à la terrestrialisation (**Fig. 8**).

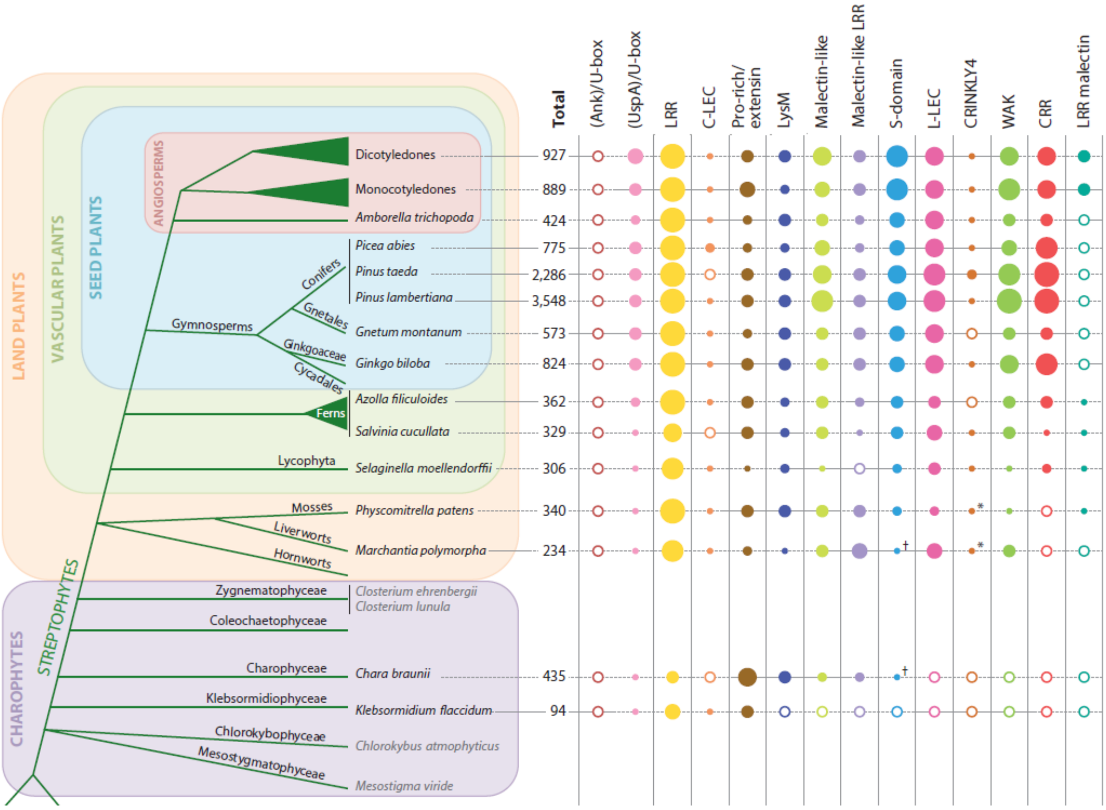


Figure 8 Présence et nombre des différentes classes de RLK (cf. article [**P28**] pour détail des classes). Les cercles vides symbolisent l'absence de gènes et la taille des cercles pleins est proportionnelle au nombre de séquences trouvées.

Ces travaux ont fait l'objet d'un article de revue publié dans Ann Rev of Plant Bot (2020) [**P28**]. Céline Gottin, que j'ai co-encadrée en thèse (Partie III.6), a contribué à l'identification des différents gènes dans l'ensemble des génomes utilisés.

III.5. Les récepteurs à LRR : vers un nouveau paradigme d'annotation

III.5.1. Contexte et objectifs

Les récepteurs kinase contenant un domaine LRR (LRR-RLK, partie III.3) appartiennent à la grande famille des récepteurs kinase, mais ne sont pas les seules protéines contenant ce domaine très particulier (**Fig. 4**). Deux autres familles de récepteurs en possèdent : (i) les **LRR-RLP**, ou 'Receptor Like Protein' qui sont également des récepteurs membranaires qui possèdent un domaine LRR et un domaine transmembranaire mais uniquement une queue cytoplasmique sans domaine kinase, et (ii) les **NLR**, ou 'Nucleotide binding site Leucine-Rich Repeat' qui sont des récepteurs intracellulaires comprenant un domaine NB-ARC et un domaine LRR (**Fig. 4**). Ces derniers sont hautement impliqués dans les résistances aux pathogènes et font l'objet de recherches très actives (Monteiro and Nishimura 2018). L'histoire évolutive de ces familles de gènes est plus que jamais au centre de l'attention étant donné la place centrale des préoccupations agroécologiques en recherche agronomique, remettant au cœur les problématiques des résistances des plantes aux stress biotiques (réduction des produits phytosanitaires) et abiotiques (évolution des systèmes de cultures pour faire face au changement climatique).

La plupart des études menées sur ces familles, au même titre que celles que nous avons faites (partie III.3 et III.4), reposent systématiquement sur l'analyse des domaines associés au domaine LRR (domaine kinase, domaine NB-ARC) (**Fig. 4**). Pourtant, la spécificité de ces 3 familles de récepteurs pour une cible étant principalement portée par le domaine LRR, **étudier l'évolution du domaine LRR et les mécanismes à l'origine de sa diversité** est d'un intérêt majeur pour comprendre leurs fonctionnements et leurs mécanismes d'adaptation. En particulier, le domaine LRR étant composé de séquences répétées, des questions spécifiques se posent, comme **l'évolution du nombre de motifs**, de **leur ordre, de la séquence même des motifs ainsi que d'éventuels échanges (ou de conversion génique) de motifs entre domaines appartenant aux différentes familles**.

Pour aborder ces questions, nous avons choisi de nous appuyer sur le riz. En effet le riz ayant été choisi comme espèce modèle des céréales dans les années 2000 (cf Partie I.1.), d'importantes ressources génomiques de qualité sont disponibles (génomes complets avec de nombreuses annotations). Ce travail était au cœur de la thèse de Céline Gottin que j'ai co-encadrée et pour laquelle nous avons obtenu un financement. Pour entreprendre les analyses du domaine LRR, un prérequis est que les trois familles de récepteurs, et en particulier leurs domaines LRR et chaque motif LRR qui les composent, soient identifiés et annotés de manière homogène, exhaustive et reproductible dans les génomes étudiés. Dans ce contexte, le travail de Céline a consisté à (1) développer un **outil d'annotation efficace et reproductible des motifs LRR dans les séquences protéiques**, (2) proposer une **méthode d'annotation génomique des récepteurs LRR** (modèle intron-exon) dans les génomes, avec une attention particulière pour les structures affectées par des **mutations non-sens** (appelées ici « non-canoniques », voir Partie III.5.3.), et (3) proposer des stratégies et bonnes pratiques pour la **diffusion**, la **reproductibilité**, la **portabilité** et la **transparence des outils et données** générés.

III.5.2. Incohérence des annotations des récepteurs à LRR chez le riz

La première partie de ce travail a consisté à construire un **pipeline d'annotation des domaines protéiques**, permettant d'identifier chaque domaine protéique, et, dans le cas du domaine LRR, les bornes précises de chaque répétition. Ce pipeline, nommé '**LRRprofiler**' reconstruit, entre autres, des profils de type 'HMM' (pour 'Hidden Markov Model') des motifs LRR de façon itérative pour chaque famille dans un protéome donné. L'outil identifie automatiquement la famille d'appartenance de chaque protéine en fonction des domaines détectés. Par rapport aux outils existants, ce pipeline s'est avéré plus efficace pour détecter les LRR contenus dans les NLR, et beaucoup plus exhaustif et précis sur la position de chaque motif. Son utilisation est extrêmement simple car il est disponible dans un 'container' de type singularity contenant toutes les briques nécessaires à son exécution et offre des sorties sous forme d'image et de tableaux facilement utilisables pour l'expertise des données.

Dans un deuxième temps nous avons été confrontés à un problème majeur, souvent mentionné mais assez rarement pris en compte dans la littérature : les **incohérences des annotations structurales de ces gènes** au niveau génomique. Sur la même version du génome de référence du riz correspondant au génotype Nipponbare, trois annotations différentes sont disponibles. Ces annotations sont issues de l'IRGSP[14], de MSU[15] et du NCB[16] et seront nommées selon ces acronymes dans le document. Une analyse comparative exhaustive a mis en évidence des **incohérences très importantes** entre elles. Notamment, **seuls 52% des locus sont détectés simultanément par les trois annotations** (**Fig. 9a**). De plus, même lorsque deux annotations prédisent un gène au même locus, ces annotations sont souvent très différentes, comme en atteste la comparaison des tailles des protéines prédites aux mêmes locus entre les annotations MSU et IRGSP (**Fig. 9b**).

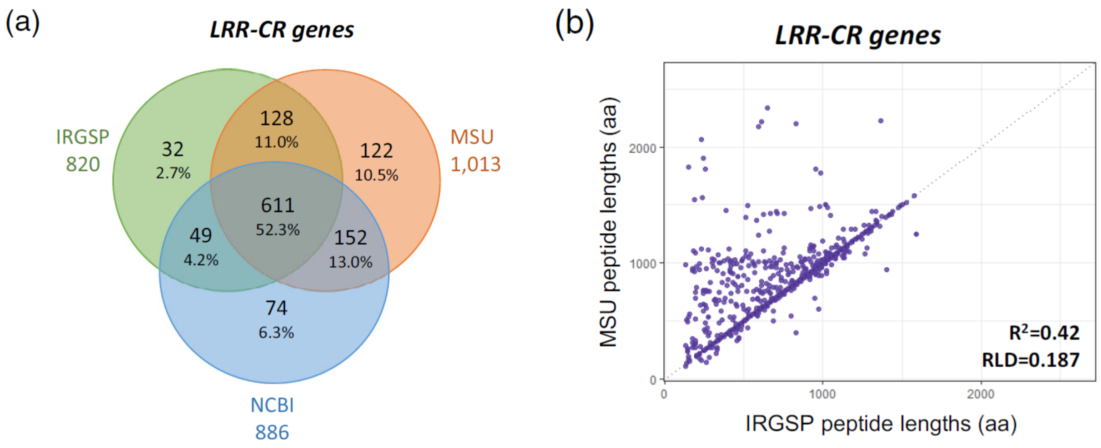


Figure 9 Comparaison des trois annotations publiques du riz ('IRGSP', 'MSU', 'NCBI') (a) diagramme de Venn représentant le nombre de modèles de gènes situés aux mêmes locus (chevauchement des annotations d'au moins une base) pour l'ensemble des récepteurs contenant un domaine LRR entre les trois annotations. (b) taille des protéines prédites par deux des annotations automatiques ('NCBI' et 'IRGSP'), chaque point représente un locus sur lequel les deux annotations ont prédit un gène [**P31**].

III.5.3. Expertise et concept de gènes 'non-canonique'

Les disparités importantes identifiées précédemment entre les différentes annotations publiques ont été un réel obstacle lorsqu'il s'est agi de constituer le jeu de données de séquences sur lequel se baser pour entreprendre des analyses évolutives du domaine LRR. Nous avons donc entrepris une **expertise** manuelle de l'annotation de ces gènes dans l'objectif de pouvoir mener à terme ces analyses comparatives. Or nous avons constaté rapidement qu'au sein des récepteurs à LRR, de **très nombreuses copies contiennent des mutations non-sens**, mais qui se sont produites suffisamment récemment pour que l'ensemble de la séquence reste quasiment intacte. Les programmes d'annotations de génomes, calibrés principalement pour identifier les gènes fonctionnels, assimilent ces copies à des gènes fonctionnels, et les annotent très souvent soit de façon erronée en les 'forçant' à coller aux contraintes de l'annotation des gènes fonctionnels, soit de façon partielle, soit pas du tout (**Fig. 10**). Dans l'exemple de la figure 10, très représentatif, seule l'annotation du NCBI identifie un pseudogène, mais les limites de la partie codante n'étant pas identifiées, il est impossible de récupérer la séquence de la protéine correspondante (telle qu'elle était avant la mutation). La version de l'annotation de l'IRGSP est, quant à elle, très partielle et celle de MSU contient un intron 'douteux' car très court et contenant une séquence présentant une forte identité avec la séquence codante d'autres copies. Ces erreurs posent de multiples problèmes. Une mauvaise annotation peut conduire à des

[14] 'International Rice Genome Sequencing Project'
[15] 'Michigan State University' dans le cadre du 'Rice Genome Annotation Project'
[16] 'National Center for Biotechnology Information'

interprétations évolutives erronées mais aussi empêcher l'identification d'allèles de résistance ou fausser les analyses de type pangénomique par exemple.

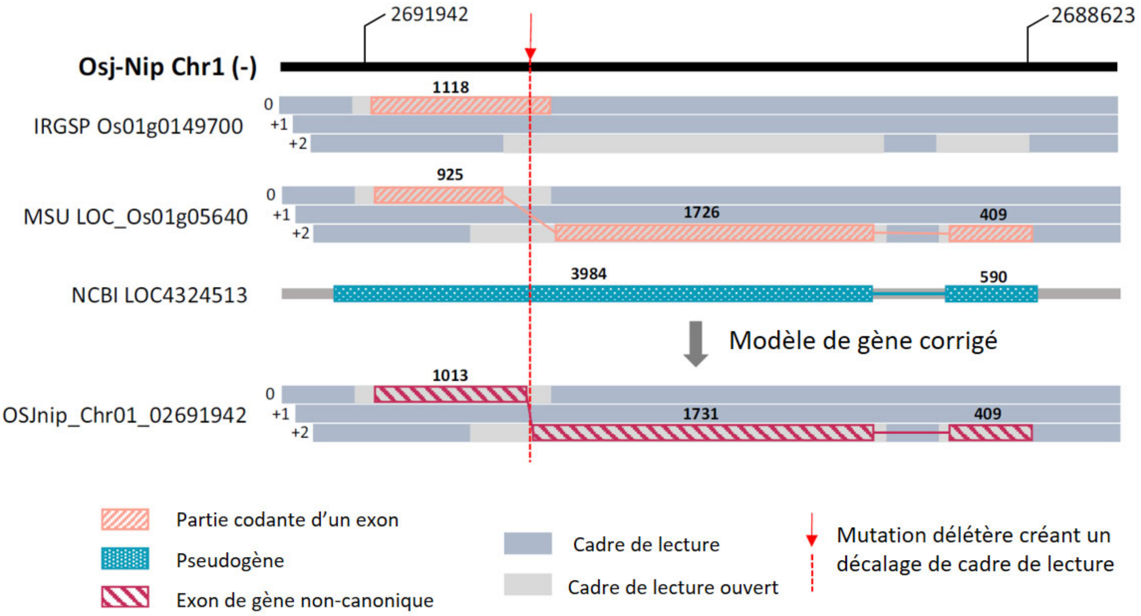


Figure 10 Représentation schématique d'un exemple d'incohérence entre les trois annotations publiques et comment l'expertise a été réalisée. Ce gène est un LRR-RLK localisé sur le chromosome 1 de Nipponbare. Les chiffres représentent la taille en paire de base. Dans cet exemple, une mutation de type indel provoque un décalage de cadre de lecture dans le premier exon du gène. L'annotation IRGSP n'a identifié que la première partie de la séquence codante, et s'arrête au premier codon stop. L'annotation MSU identifie une séquence codante plus longue mais met de côté la mutation en introduisant un intron 'douteux' pour récupérer le bon cadre de lecture. L'annotation du NCBI identifie un pseudogène, mais la séquence protéique ne peut pas être déduite. Dans la correction, le cDNA recouvre l'ensemble de la séquence codante initiale, dans les bons cadres de lectures grâce à l'identification de la mutation. Le gène est marqué 'non-canonique' et la mutation repérée.

Suite à ces constatations, nous avons choisi de mener une **expertise exhaustive de l'ensemble des gènes** avec l'objectif de récupérer la totalité des séquences des gènes codant pour les récepteurs à LRR, même s'ils contiennent des mutations non-sens. Un point essentiel a donc été de proposer **une nouvelle façon d'appréhender ces copies de gènes contenant des mutations non-sens**. Nous avons ainsi défini la catégorie de **gènes non-canoniques**, qui sont des copies de gènes dans lesquels nous pouvons observer au moins un des éléments suivants : (1) présence d'une mutation responsable de l'apparition d'un codon stop (2) présence d'une mutation de type indel provoquant un décalage de phase, (3) perte (par indel ou mutation) du codon start ou du codon stop ou (4) absence de sites d'épissage GT-AG ou GC-AG. Pour cela ont été combinées les informations au niveau protéique (présence des domaines, ordre, longueurs, espacement) et génomique (structure intron/exon attendue, présence d'ORF à proximité, introns 'douteux', sites d'épissage). Nous avons identifié **environ 30% de copies non-canoniques,** i.e. contenant des mutations non-sens (**Tableau 2**).

Tableau 2 Nombre de gènes identifiés de chaque famille dans les différentes annotations. Les annotations publiques sont IRGSP, MSU, NCBI et notre annotation expertisée est mentionnée par 'exp.'. Les pourcentages des lignes 1 à 4 sont calculés par rapport au total (colonne 1) et ceux des lignes 5 et 6 par rapport au nombre de gène de chaque famille (ligne 4).

	total	LRR-RLK	LRR-RLP	NLR	LRR-CR non classés
IRGSP	1047	237 (23%)	160 (15%)	282 (30%)	368 (35%)
MSU	1226	329 (27%)	141 (12%)	418 (34%)	338 (28%)
NCBI	1073	305 (28%)	121 (11%)	361 (34%)	286 (27%)
exp. total	1058	350 (33%)	147 (14%)	503 (48%)	58 (5%)
exp. Canonique	784 (74%)	297 (85%)	99 (67%)	320 (64%)	36 (62%)
exp. Non-canonique	274 (26%)	53 (15%)	48 (33%)	183 (36%)	22 (38%)

Au final, l'ensemble de cette expertise a permis d'identifier 1058 gènes, répartis dans différentes sous-familles (**Tableau 2**). Parmi eux, 328 (soit 31%) ont été modifiés au cours de l'expertise car aucune des trois annotations publiques n'était satisfaisante. Si ce nombre total est similaire à celui des annotations automatiques, l'annotation, plus complète, permet une bien meilleure assignation automatique dans les sous-familles.

Le choix conceptuel que nous avons fait, *i.e.* annoter des gènes dans lesquels nous avons identifié des mutations non-sens et qualifier ces gènes de non-canoniques, et non de 'pseudogènes', est essentiel. En effet, cette annotation est purement factuelle et ne contient pas d'interprétation d'ordre fonctionnel. En contrepartie, l'ensemble de l'information est disponible, visualisable et les séquences disponibles dans différents formats. Les gènes non-canoniques représentent une vaste diversité de cas, certains ne permettant vraisemblablement pas la traduction d'une protéine fonctionnelle et d'autres pouvant très bien être fonctionnels. Par exemple, l'apparition d'un codon stop peut avoir un impact très différent en fonction de sa position plus ou moins en amont du codon stop initial. De même, si le codon stop est perdu, plusieurs acides aminés seront ajoutés, dont le nombre dépendra de la position du prochain codon stop rencontré dans la séquence. Prédire quelle sera la conséquence de cette modification n'est pas possible *in silico*.

III.5.4. Transfert d'annotation et comparaison de génotype

Pour pouvoir exporter cet effort d'annotation vers d'autres génomes et faire bénéficier la communauté de cette nouvelle approche, un **pipeline de transfert d'annotation** appelé **'LRRtransfer'** a été construit. Ce pipeline combine une étape de définition de zones cibles en partant des gènes expertisés, et une étape de transfert d'annotation proprement dit selon trois méthodes en fonction du degré d'homologie entre le gène expertisé et la zone cible à annoter. Ce pipeline a été utilisé sur un deuxième génotype de riz, **Kitaake**. L'utilisation de ce pipeline, suivi d'une étape d'expertise beaucoup plus réduite, a permis de fournir une annotation de même qualité sur Kitaake. 1064 gènes contenant des LRR ont été identifiés, dont 1010 appartiennent aux trois sous-familles étudiées (**Fig. 11**). **114 nouveaux locus ont été identifiés** (locus non présent dans l'annotation publique de Kitaake) dont 48 sont canoniques.

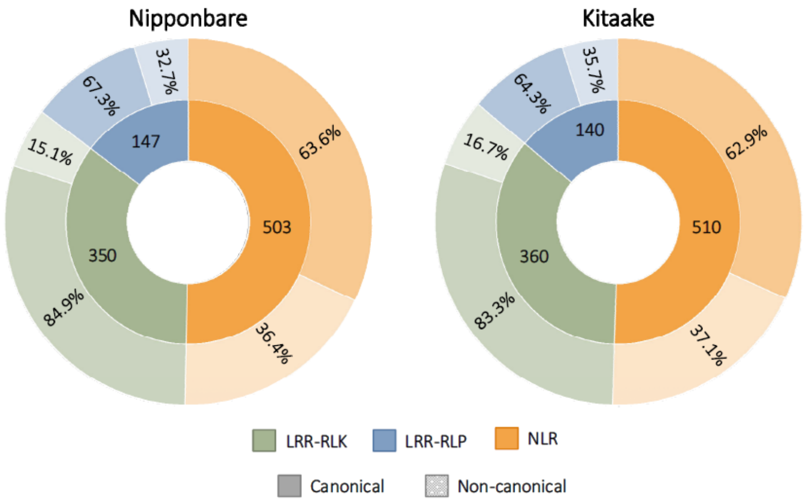


Figure 11 Proportion de gènes canoniques et non canoniques par sous-familles dans les annotations expertisées de Nipponbare et Kitaake. Les pourcentages sont calculés par sous-famille dans le cercle extérieur.

La **comparaison des deux génotypes** a permis d'apporter de premiers éléments de réponse aux questions relatives à la **diversité et à la variabilité du contenu en gènes pour ces récepteurs**. Globalement, le nombre de gènes est équivalent entre les deux génotypes, leur répartition en sous-familles et les proportions de gènes non canoniques également (**Fig. 11**). La recherche des paires allèliques entre les deux génotypes a permis d'identifier 1002 paires d'allèles. Ces allèles ont été comparés, en particulier leur nombre de motifs LRR au sein du domaine LRR (**Fig 12**). La corrélation entre le nombre de motifs LRR des allèles de chaque génotype est beaucoup plus forte au sein de l'annotation expertisée (R^2=0.98 avec 0.5 motifs de différence en moyenne) que pour l'annotation publique (R^2=0.67 avec 2.8 motifs de différence en moyenne). Ce résultat démontre le gain en qualité de l'annotation expertisée. En effet, le niveau d'identité nucléotidique entre les deux génotypes oscille autour de 98% en moyenne sur l'ensemble du génome, et il est attendu que le nombre de LRR entre les allèles soit très conservé (Jain et al. 2019).

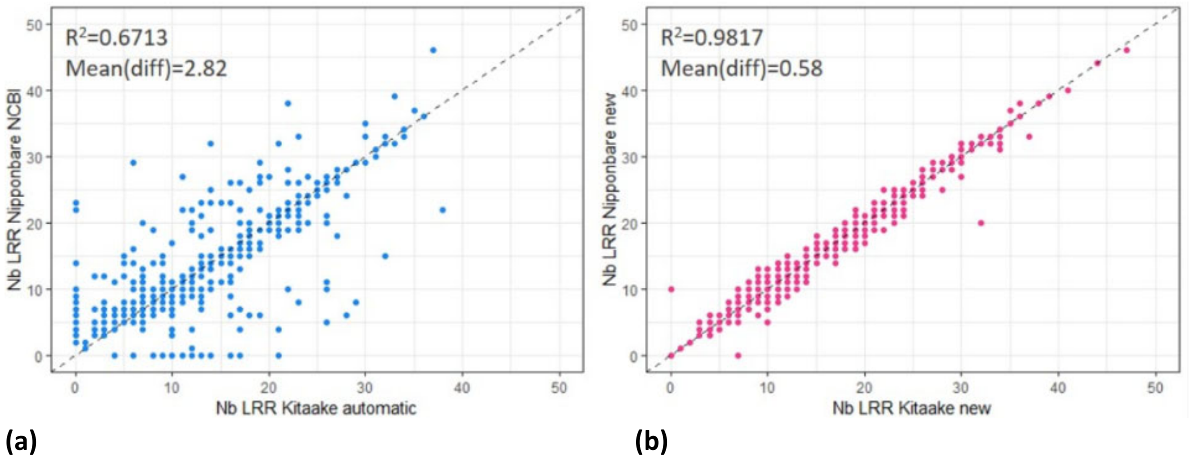


Figure 12 Nombre de motifs LRR entre les allèles identifiés dans les deux génotypes Nipponbare et Kitaake, chaque point représentant un locus. (a) résultat obtenu avec les annotations automatiques (b) graphique obtenu avec nos annotations expertisées. Les motifs sont détectés sur les protéines prédites avec le pipeline LRRProfiler.

Ces comparaisons permettent de mettre en évidence **48 paires pour lesquelles un gène est canonique chez l'un des génotypes et non canonique chez l'autre**. Ces cas sont particulièrement intéressants car ils permettent de repérer une partie de la variabilité qui peut avoir des **conséquences phénotypiques**. Il est notable que parmi ces 48 paires, plus de la moitié (27) concerne des NLR, principalement impliqués dans les résistances.

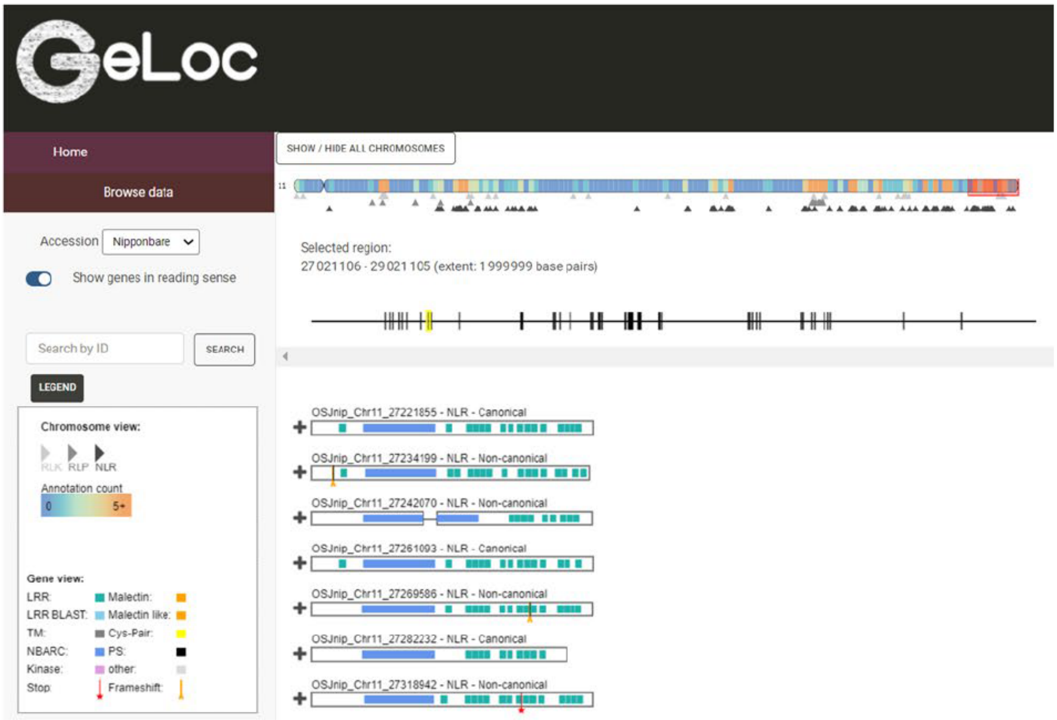
Enfin l'analyse des **PAV** ('Presence Absence Variations') entre les deux génotypes a mis en évidence 48 gènes présents uniquement chez Nipponbare et 58 uniquement chez Kitaake. La moitié de ces gènes sont assez divergents des autres copies du génome (moins de 80% d'identité protéique). Certains de ces gènes (5 chez Kitaake et 7 chez Nipponbare) montrent, en revanche, une similarité protéique élevée (plus de 95%) avec des protéines issues de cultivars de type *indica* (l'autre sous-espèce des riz asiatiques cultivés) questionnant sur d'éventuelles introgressions résultants de flux de gènes ou des programmes de sélection. Il est notable également que 40% de ces gènes sont non canoniques. Ces résultats sont cohérents avec différentes études pangénomiques qui montrent que les récepteurs à LRR sont particulièrement représentés dans les génomes accessoires (Zhang et al. 2016; Dolatabadian et al. 2017). De plus ces variations de type PAV peuvent être à l'origine d'une variabilité phénotypique sur des traits liés aux fonctions telles que les signalisations cellulaires en réponse à des stress biotiques et abiotiques.

L'ensemble de ce travail constitue la thèse de Céline Gottin et a été publié [**P31**].

III.5.5. Valorisation et conclusion

Une collaboration avec Marilyne Summo (AGAP, ingénieur du plateau de bioinformatique) nous a permis de permis d'élaborer un site de visualisation des données intégrant l'ensemble des résultats cités et sur lequel l'ensemble des données peut être téléchargé (**Fig. 13**). Le site est disponible sur le site suivant : https://rice-genome-hub.southgreen.fr/content/geloc.

Ce travail a mis en évidence le **manque de fiabilité des données issues des procédures automatiques d'annotation lorsqu'il s'agit d'annoter des familles de gènes complexes**. Les procédures automatiques gèrent de façon hétérogène, et parfois de façon erronée, les copies de gènes qui présentent des mutations non-sens, pour lesquelles le processus de non fonctionnalisation est peut-être enclenché. Nous avons choisi cette terminologie '**non-canonique**', au lieu de 'pseudogène' à dessein, pour ne **pas faire d'inférence sur la potentielle fonctionnalité** de ces copies de gènes contenant des mutations non-sens et pouvoir les inclure dans les jeux de données. Les comparaisons du nombre de motifs LRR des gènes mettent en évidence l'impossibilité d'étudier l'évolution du domaine LRR à partir des annotations automatiques (**Fig. 12**). Il serait définitivement hasardeux de baser des analyses évolutives du domaine LRR (nombre de motifs, séquences, conversion…) sur des annotations automatiques.



Accession Nipponbare

Show genes in reading sense

Search by ID

Selected region:
27 021 106 - 29 021 105 (extent: 1 999999 base pairs)

OSJnip_Chr11_27221855 - NLR - Canonical
OSJnip_Chr11_27234199 - NLR - Non-canonical
OSJnip_Chr11_27242070 - NLR - Non-canonical
OSJnip_Chr11_27261093 - NLR - Canonical
OSJnip_Chr11_27269586 - NLR - Non-canonical
OSJnip_Chr11_27282232 - NLR - Canonical
OSJnip_Chr11_27318942 - NLR - Non-canonical

Figure 13 Visualisation des données de l’annotation expertisée des récepteurs contenant des LRR, exemple de Nipponbare. Vue **globale d’un chromosome** dont une zone peut être sélectionnée pour obtenir une visualisation détaillée de la **structure des gènes d’intérêt** qu’elle contient. La visualisation de chaque gène intègre la **structure génomique** (intron/exon), le sens de lecture, la position d’éventuels codon stop prématurés ou de changement de phase de lecture, et la structure **protéique** (position des domaines fonctionnels).

Le développement d’un outil de transfert d’annotation (**LRRtransfer)** permet d’envisager, de façon réaliste, de procéder à une annotation fiable de l’ensemble des récepteurs à LRR, incluant les gènes non canoniques, dans plusieurs autres génomes de riz. Moyennant quelques développements et adaptation du pileline, une telle annotation serait également envisageable dans d’autres espèces. L’outil **LRRprofiler** permet de déterminer précisément le nombre et les positions des LRR dans les protéines ainsi prédites. Ces avancées permettent d’envisager d’aller plus loin dans l’étude de la diversité de ces familles de récepteurs (variabilité allèlique, canonique vs. non canonique au sein du ‘core-genome’ et du genome ‘accessoire’) et dans l’étude de l’évolution du domaine LRR chez les plantes (cf. partie ‘Projet de recherche’).

IV. Epissage alternatif

IV.1. Pourquoi s’intéresser au mécanisme de l’épissage alternatif ?

Il y a plusieurs niveaux de motivation. Le premier est que ce mécanisme est assez fascinant pour avoir remis en question le ‘dogme’ : un gène – une protéine. L’épissage alternatif (‘Alternative Splicing’) est le mécanisme par **lequel plusieurs molécules d’ARN messager (ARNm) différentes (appelées isoformes) résultent de la maturation de la même molécule précurseur d’ARNm** transcrite initialement (Nilsen and Graveley 2010) (**Fig. 14**). En faisant qu’un gène code pour plusieurs protéines différentes, l’épissage alternatif offre un niveau additionnel de complexité entre le contenu en gène d’un individu et son phénotype. Ce mécanisme fait l’objet d’une régulation fine à l’échelle des différents organes et des stades de développement d'un organisme et est, de surcroit, très sensible aux stress environnementaux. Les toutes premières découvertes de ce mécanisme datent des années 70-80 ; depuis, sa place prépondérante dans l'ensemble des eucaryotes a été démontrée (Irimia and Roy 2014).

Un autre niveau de motivation est son lien inattendu avec les duplications de gènes et les familles multigéniques évoquées précédemment. En effet, il a été montré qu'il existe une corrélation négative entre la taille d'une famille multigénique est la propension qu'ont les gènes à présenter de l'épissage alternatif (Kopelman et al. 2005). L'épissage alternatif a été décrit comme le moyen, pour un gène donné, de disposer d'une sorte de 'paralogue intrinsèque'. Ces deux mécanismes sont des contributeurs majeurs de la diversification protéique d'un organisme, et les relations entre les deux commencent à être explorées, montrant notamment que les gènes dupliqués ont des profils d'épissage alternatif différents (Tack et al. 2014; Iniguez and Hernandez 2017).

La troisième raison est plus pragmatique et tient au fait que lorsque l'on étudie le contenu en transcrit ou en protéine d'un génotype il faut décider comment gérer les différentes isoformes existantes pour un même gène. Une possibilité est de choisir 'le plus long' ou 'le plus exprimé', si l'information est disponible. Ceci ne pose pas de réels problèmes pour une analyse globale de polymorphisme ou pour la recherche de SNP. Cependant, se confronter à la réalité biologique et aborder les nouvelles questions qu'elle suscite est un des moteurs de nos métiers.

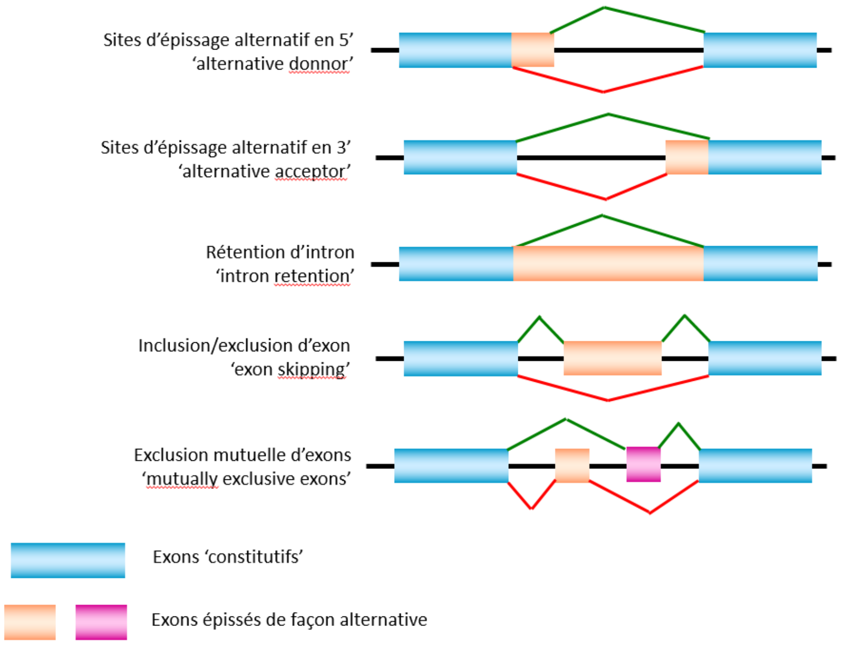

Sites d'épissage alternatif en 5'
'alternative donnor'
Sites d'épissage alternatif en 3'
'alternative acceptor'
Rétention d'intron
'intron retention'
Inclusion/exclusion d'exon
'exon skipping'
Exclusion mutuelle d'exons
'mutually exclusive exons'
Exons 'constitutifs'
Exons épissés de façon alternative

Figure 14 Différents types d'évènements d'épissage alternatif. Les boîtes de couleur représentent les exons et les lignes noires horizontales les introns. Les lignes de couleur rouge et verte représentent les connexions entre les exons après épissage, mettant en évidence les deux isoformes issues de chaque évènement, sauf pour l'exemple de la rétention d'intron pour laquelle l'une des deux formes sera sans intron.

IV.2. Mécanisme moléculaire et état des connaissances chez les plantes

Une des étapes de la maturation des molécules d'ARNm précurseurs est le processus d'épissage, pendant lequel les introns sont excisés grâce à un complexe protéique appelé 'spliceosome', dont l'action est modulée par un ensemble d'activateurs agissant en *trans* (protéines ou ARN additionnels) et en *cis* (appartenant à la molécule d'ARNm elle-même). Détecté chez tous les eukaryotes, l'épissage alternatif a été largement étudié chez l'homme (Modrek and Lee 2002) et d'autres animaux (Barbosa-Morais et al. 2012). L'épissage alternatif augmente la complexité des transcriptomes, ce qui a deux principales conséquences : la complexification du protéome lui-même et l'ajout d'un niveau de régulation lié à l'interférence des formes alternatives sur la cinétique de dégradation des ARNm via un mécanisme appelé 'NMD' pour 'nonsense-mediated decay' ou dégradation des ARNm non-sens. La dégradation des ARNm non-sens est un processus moléculaire durant lequel les isoformes d'ARNm possédant un codon stop prématuré sont dégradés (Lewis et al. 2003; McGlincy and Smith 2008). L'épissage alternatif a été évoqué comme une des explications de l'ampleur des différences phénotypiques entre des espèces qui partagent pourtant un répertoire de gènes assez similaire comme les vertébrés (Barbosa-Morais et al. 2012).

L'épissage alternatif est donc reconnu comme une étape clé entre la transcription et la traduction (Kornblihtt et al. 2013). Pour un même individu, il varie d'un organe à un autre, selon le stade de développement et même selon le type cellulaire (Lopez-Diez et al. 2013). Sa régulation est fine et complexe et son bon déroulement est l'assurance d'un développement 'conforme' pour un organisme donné (Chen and Manley 2009). De nombreuses maladies génétiques chez l'homme ont d'ailleurs été démontrées comme étant causées par le dérèglement du processus d'épissage alternatif d'un ou de plusieurs gènes (Caceres and Kornblihtt 2002; Cieply and Carstens 2015).

L'étude de l'épissage alternatif chez les plantes est plus récente (Kazan 2003) et a été réalisée à l'échelle du génome entier d'abord chez les espèces modèles. La proportion de gènes, contenant des introns, affectés par l'épissage alternatif atteint 61% chez Arabidopsis (Marquez et al. 2012) et 48% chez le riz (Lu et al. 2010). Maintenant que les technologies de séquençage type RNAseq sont plus accessibles, l'épissage alternatif est décrit dans de nombreuses espèces : *Brachypodium distachyon* (Walters et al. 2013), la vigne (Potenza et al. 2015), l'orge (Panahi et al. 2015) ou la tomate (Sun and Xiao 2015) pour n'en citer que quelques-unes.

Malgré l'ensemble de ces recherches, il reste de nombreuses questions sur l'épissage alternatif. En particulier, la proportion d'ARNm qui est réellement traduite en protéines fonctionnelles n'est pas vraiment connue (Ezkurdia et al. 2012; Reixachs-Sole and Eyras 2022), et l'impact réel de l'épissage alternatif sur la diversification des protéines est une question encore débattue (Severing et al. 2009) (Blencowe 2017; Chaudhary et al. 2019). Mais même si toutes les isoformes d'ARNm ne sont pas destinées à être traduites, l'épissage alternatif est pressenti comme ayant un rôle fondamental dans la régulation de l'expression des gènes, notamment via la dégradation des ARNm non-sens. Par ailleurs, la régulation de l'épissage alternatif est très sensible aux conditions environnementales. Il a été montré que les profils d'épissage alternatifs changent fortement chez les plantes en réponse à des stress environnementaux (pour quelques revues de synthèse voir (Staiger and Brown 2013; Ding et al. 2014; Filichkin et al. 2015)). Des études récentes mettent de plus en plus souvent en évidence le rôle du stress dans la modification des profils d'épissage alternatif, comme des stress salins (Feng et al. 2015; Panahi et al. 2015) ou hydrique (Thatcher et al. 2016). Les conséquences moléculaires de ces modifications sont néanmoins encore mal connues mais ces dernières années, de plus en plus d'éléments mettant en lumière le rôle de l'épissage alternatif dans l'adaptation ont été mis à jour (Bartok et al. 2013; Verta and Jacobs 2022).

IV.3. Epissage alternatif et adaptation

L'une des premières questions que nous nous sommes posées concernait le **rôle potentiel de l'épissage alternatif dans l'adaptation**. Nous avons **utilisé la domestication comme un exemple de processus d'adaptation et étudié si, et comment, la domestication a affecté l'épissage alternatif**. En effet, dans le cadre de la réflexion collective portée par le sous-projet 1 du projet ARCAD (**Encadré 2**), plusieurs questions relatives à l'impact de la domestication ont été abordées. Les transcriptomes de dix accessions appartenant aux compartiments sauvages et cultivés (accompagnés de deux groupes externes), ont été séquencés pour une dizaine d'espèces jalonnant la phylogénie des angiospermes. Les données de transcriptomique obtenues dans le cadre de ce projet représentaient une opportunité intéressante pour étudier si la domestication a eu un impact sur le niveau d'épissage alternatif global ou sur celui de certains gènes. Nous avons choisi de réaliser cette étude sur le sorgho. L'ensemble de ce travail, auquel a grandement contribué Audrey Serra, étudiante en césure de l'INSA de Lyon que j'ai co-encadrée, a fait l'objet d'une collaboration avec David Pot (équipe GIV AGAP) [**P24**].

La première étape de cette étude a consisté à réaliser le 'mapping' des fragments séquencés ('reads') sur la séquence génomique complète du sorgho, puis d'utiliser différents outils disponibles dans la littérature pour interpréter le mapping, pour chaque gène, en termes de fréquence relative des différentes isoformes, le cas échéant. Nous avons appliqué un seuil de détection empirique d'une couverture moyenne de 5 reads pour déclarer une isoforme. Nous avons enfin sélectionné les gènes pour lesquels deux isoformes, et deux seulement, étaient exprimées au moins chez un individu sauvage et un individu cultivé. Au final, notre jeu de donnée comportait environ **1300 gènes** pour lesquels nous avons résumé l'information du niveau d'expression des **deux isoformes** sous la forme d'un **ratio de niveau d'expression** pour mesurer **l'équilibre entre les deux isoformes.** Nous avons fait **le rapport entre le niveau d'expression de l'isoforme la plus exprimée** (en moyenne) **sur le niveau total d'expression du gène**.

Les résultats montrent une **réduction de la variabilité du ratio entre deux isoformes dans les formes cultivées par rapport aux formes sauvages**. Sur le graphique de la **Fig. 15**, les variances de ces ratios calculés au sein des individus sauvages sont plus élevées que les variances des ratios obtenues au sein des individus cultivés. Ces différences se sont avérées significatives quel que soit le sous-échantillonnage de 6 individus considéré. Ce résultat fait écho aux résultats classiquement observés lorsque l'on compare la diversité nucléotidique des formes sauvage et cultivée d'une même espèce : les formes cultivées présentent en général un niveau de diversité réduit par rapport aux formes sauvages, résultat de l'effet d'échantillonnage lié à la domestication (seule une partie de la diversité totale de l'espèce est à l'origine de la forme cultivée).

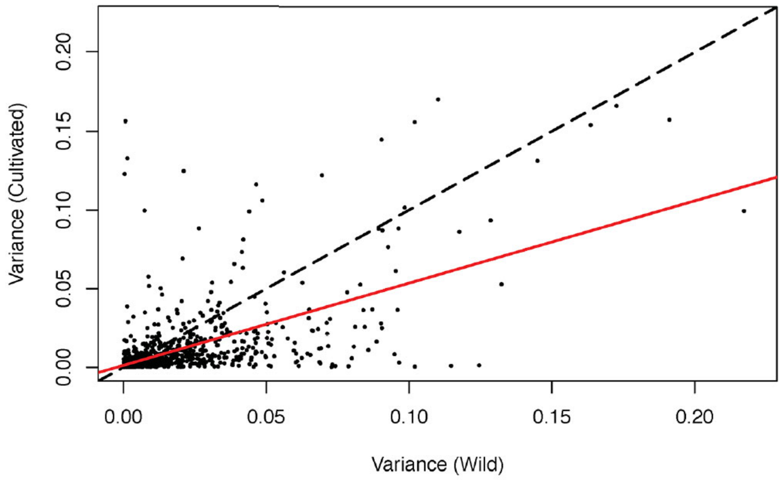


Figure 15 Comparaison des variances des ratios d'expressions (entre les deux isoformes de chaque gène, représenté par un point) au sein de l'échantillon sauvage (abscisse) et cultivé (ordonnée). La droite en rouge est la régression linéaire de ces points (pente 0.54) alors que la droite en pointillés est la diagonale y=x.

Nous avons examiné deux types de gènes : (1) des gènes ayant les écarts les plus importants entre les formes sauvages et cultivées comme par exemple les gènes pour lesquels les ratios d'expression entre les deux isoformes étaient très similaires au sein des individus cultivés mais très variables au sein des individus sauvages et (2) les gènes pour lesquels la moyenne des ratios était très différente entre les formes sauvages et cultivées, i.e. ceux pour lesquels l'isoforme majoritaire était différente entre les compartiments. Certains de ces gènes présentaient une identité forte avec des gènes impliqués dans le contrôle de traits relatifs au syndrome de domestication ou dont le terme GO était intitulé 'régulation de la qualité biologique'. D'autres sont de fonction inconnue, mais pourraient être des candidats intéressants pour des analyses fonctionnelles (via la recherche de mutants par exemple). En effet ces différences d'expressions entre isoformes pourraient permettre d'identifier des gènes de domestication qui auraient pu échapper aux méthodes classiques de recherches de traces de sélection ou de génétique quantitative (QTL/GWAS).

Cette étude a montré que **la domestication a modifié les patrons d'épissage alternatif**. Il est probable que la perte de diversité nucléotidique classiquement associée à la domestication ait affecté des séquences régulatrices contrôlant la transcription et/ou la dégradation des isoformes. Si l'impact de la domestication sur l'expression des gènes commençait à être décrit (Rapp et al. 2010; Bellucci et al. 2014; Liu et al. 2019), notre étude était l'une des premières à montrer l'impact de la domestication sur l'équilibre entre des isoformes. Depuis, d'autres auteurs ont montré des différences de profils d'épissage alternatif entre formes cultivées et sauvages, sur le tournesol notamment, et suggéré que la domestication ait pu entrainer la sélection d'allèles de régulation aux effets pléiotropiques (Smith et al. 2018). Un effet de la domestication, spécifiquement sur l'épissage alternatif, a aussi été mis en évidence chez le blé (Yu et al. 2020). Sur l'expression des gènes en général, une étude récente sur l'olivier a suggéré que la domestication n'a eu que des conséquences génomiques modérées mais que le syndrome de domestication est principalement lié à des changements dans l'expression des gènes

(Gros-Balthazard et al. 2019). Sur le piment, d'autres auteurs ont émis l'hypothèse que des changements de profils d'expression aient pu être directement causés par les pressions de sélection au cours de la domestication (Martinez et al. 2021). À l'instar des données de polymorphisme, les données d'expression, incluant les profils d'épissage, peuvent ainsi être utilisées pour identifier d'autres types de gènes impliqués dans le contrôle de caractères liés au syndrome de domestication.

PROJET DE RECHERCHE

Dans la suite de mes travaux, je souhaite approfondir les questionnements à l'interface entre, structure et fonctionnement moléculaire du génome, et biologie évolutive, dans deux thématiques : (1) l'étude de l'évolution des récepteurs à LRR et en particulier du domaine LRR, et (2) l'étude de l'évolution et du rôle de l'épissage alternatif dans l'adaptation.

I. Evolution des récepteurs à LRR

Aujourd'hui, le monde agricole et les acteurs de la recherche sont confrontés à deux enjeux majeurs : faire face aux **changements climatiques** et **diminuer l'apport d'intrants, notamment des produits phytosanitaires, dans les agrosystèmes**. Ces deux enjeux ne sont pas indépendants, notamment parce que les changements climatiques ont des impacts importants sur l'aire de répartition des agents pathogènes qui a tendance à s'étendre vers les pôles (Bebber et al. 2013). Les attaques des plantes par les agents pathogènes sont toujours un problème majeur en agriculture (Savary et al. 2019). Elles provoquent des pertes de productivité, menacent la sécurité alimentaire et sont à l'origine de l'utilisation massive de produits phytosanitaires. L'instabilité climatique, quant à elle, nous enjoint à étudier, et à utiliser, les capacités des plantes à percevoir et faire face aux stress abiotiques (sècheresse, salinité, ressources limitées …) pour promouvoir une agriculture plus résiliente. Proposer des variétés, des espèces ou des mélanges plus adaptés à des environnements changeants ou agressants, est une des stratégies susceptibles d'aider l'agriculture à surmonter ces défis. Mieux connaitre le fonctionnement biologique des plantes face à ces stress et les gènes sous-jacents peut contribuer à guider la construction de nouveaux idéotypes. Parmi les mécanismes de signalisation cellulaires, essentiels pour le bon fonctionnement de la plante, les **récepteurs à LRR jouent un rôle biologique majeur** en intervenant en amont, dans la **perception de l'environnement**, puis dans le déploiement du signal et des réponses cellulaires de la plante.

Pour rappel, les **récepteurs à LRR** des plantes (LRR-RLK, LRR-RLP et NLR, **Fig. 4** , partie III.3 et III.5) sont centraux dans les mécanismes de **perception de l'environnement**. De nombreuses fonctions ont été attribuées aux LRR-RLK et LRR-RLP comme la croissance et le développement (traits affectés en réponse à des stress environnementaux) mais aussi la symbiose et **l'immunité**. Les LRR-RLP ont principalement besoin d'interagir avec des corécepteurs de type LRR-RLK pour assumer leur fonction. Les NLR quant à eux sont principalement impliqués dans l'immunité (Barragan and Weigel 2021). La diversité génétique et fonctionnelle de l'immunité des plantes réside notamment dans ce vaste répertoire de récepteurs (Pilet-Nayel et al. 2017; Kourelis and van der Hoorn 2018; Nelson et al. 2018). La **diversité intra- et interspécifique** de ces récepteurs est une **ressource potentiellement exploitable pour l'amélioration des plantes et la protection durable des cultures**. Néanmoins, pour réussir à utiliser ces ressources, il faut **documenter précisément cette diversité, comprendre comment elle est générée et quelle est sa dynamique évolutive**. Enfin, ces récepteurs partagent le **domaine LRR**, particulièrement variable, qui est au cœur des mécanismes de reconnaissance protéine-protéine (**Fig. 4**).

Les questions que je souhaite traiter dans les prochaines années s'inscrivent dans la continuité des recherches initiées sur ces familles de récepteurs. En particulier, il s'agira de déterminer **quel est le réel niveau de diversité intraspecifique de ces récepteurs** ? Comment s'est construite cette diversité intra- et interspécifique et **quel rôle la sélection a-t-elle joué** ? **Comment le domaine LRR évolue** (mutations ponctuelles, duplication de motifs, recombinaison, conversion génique, …) à l'échelle intra-et interspécifique ? L'objectif, à terme, étant de pouvoir mobiliser ces ressources génétiques, et les connaissances qui y seront associées, pour participer soit à la construction d'idéotypes (utilisation d'allèles existants, dans des variétés ou des mélanges, recours à la variabilité des espèces apparentées aux espèces cultivées ou 'Crop Wild Relatives'), soit à la conception de nouveau allèles (modification du domaine LRR par recombinaison ou édition par exemple (Veillet et al. 2020 ; Guyon-Debast et al. 2021)).

Les espèces sur lesquelles je propose de m'appuyer pour traiter ces questions sont (1) le riz et ses apparentées sauvages et (2) les espèces diploïdes apparentées aux blés cultivés (blé dur). En effet, à l'heure actuelle, le riz reste la céréale pour laquelle les ressources moléculaires sont les plus nombreuses. Néanmoins, à plus long terme, mon souhait serait de transférer sur le blé les connaissances et méthodologies acquises sur le riz. Pour cela, une possibilité est d'utiliser les espèces de blés diploïdes (genres *Triticum* et *Aegilops*) comme intermédiaire, pour s'affranchir, dans un premier temps, des difficultés liées à la polyploïdie.

I.1. Annotation expertisée

Récemment, un important projet de séquençage *de novo* de haute qualité (qualifié de PSRefSeq pour 'Platinium Standard Reference Sequences') de 16 génotypes de riz jalonnant la diversité de l'espèce a été entrepris (Zhou et al. 2020). Ces génotypes représentent les 15 sous-populations du riz asiatique cultivé (Mussurova et al. 2020; Zhou et al. 2020) (**Fig. 16a**). Cet effort offre à la communauté internationale une ressource précieuse, que je propose d'utiliser pour la caractérisation génétique des récepteurs à LRR. Par ailleurs, plusieurs génomes complets sont disponibles au niveau interspécifique dans ce complexe d'espèces (Stein et al. 2018) dont la phylogénie est connue (Wambugu et al. 2015) (**Fig. 16b**).

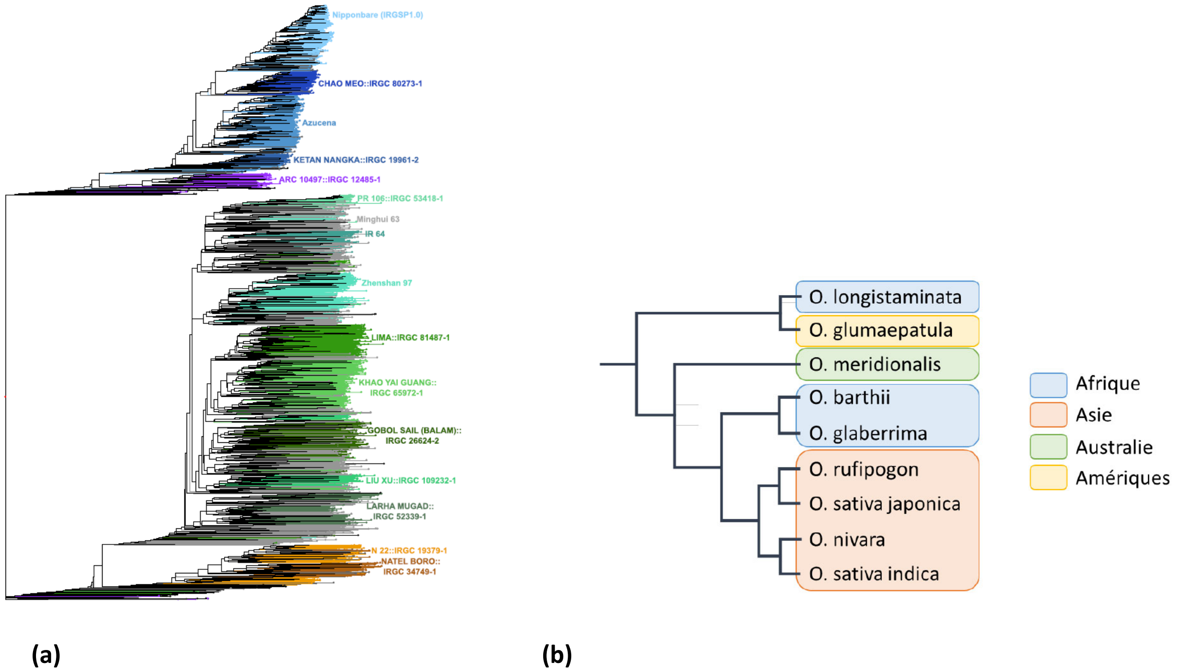


Figure 16 Ressources génomiques au sein des riz. (a) Position des accessions séquencées selon le standard PSRefSeq (Zhou et al. 2020) au sein de la variabilité génétique intraspécifique. (b) Phylogénie des espèces de riz de génome A, adaptée de (Wambugu et al. 2015).

Comme nous l'avons montré au cours de la thèse de Céline Gottin [**P31**], les récepteurs à LRR sont souvent très mal annotés à cause de leur organisation génomique et de leur diversité (partie III.6). Pour pouvoir étudier l'évolution de ces récepteurs la première étape incontournable sera de **réaliser une annotation fiable de ces récepteurs**. Je propose d'utiliser sur ces génomes de haute qualité les méthodes et concepts que nous avons développés (LRRtransfer, LRRprofiler), complétés par des outils disponibles comme NLR-Annotator (Steuernagel et al. 2020).

Parallèlement il sera primordial de veiller à ce que l'ensemble des données produites soient accessibles, traçables et visualisables. La stratégie sera d'intégrer l'ensemble des données produites sur le site GeLoc développé avec Marilyne Summo et d'enrichir ce site de nouvelles fonctionnalités pour faciliter les études comparatives.

I.2. Pangénome et série d'orthologues des récepteurs à LRR : étude de l'évolution des récepteurs et du domaine LRR

Dans un deuxième temps, la comparaison du répertoire des récepteurs entre les différents génotypes nous permettra de construire leur pan-LRRome et de générer une matrice de présence absence ('PAV') (Barragan and Weigel 2021). Les pangénomes sont définis, au sein d'un échantillons d'individus d'une même espèce, comme la combinaison du 'core' genome qui contient les gènes présents chez tous les individus considérés, et le génome accessoire ou 'dispensable' composé des gènes qui sont absents de certains individus (Tranchant-Dubreuil et al. 2019; Bayer et al. 2020). Un enrichissement en récepteurs de type NBS-LRR a déjà été montré dans le pangénome construit chez le riz notamment (Zhao et al. 2018). Dans cette étape, une attention particulière sera portée aux clusters de gènes, zones dans lesquelles plusieurs gènes sont physiquement proches les uns des autres. Ces clusters, qui contiennent souvent des gènes canoniques et non canoniques, sont particulièrement enclins aux duplications et présentent souvent un niveau élevé de diversité qui se maintient parfois sur de longues périodes évolutives, au-delà des évènements de spéciation (Mizuno et al. 2020). À partir de ces données il s'agira (1) de rechercher l'origine des variations de présence/absence pour les différentes familles de récepteurs à LRR et de déterminer comment les duplications et pertes de gènes ont façonné cette diversité et (2) de rechercher des traces de sélection dans la diversité allélique.

Au niveau interspécifique, l'approche sera d'identifier des séries d'orthologues entre les espèces. Des signatures de sélection seront recherchées dans la variabilité interspécifique en utilisant les méthodes contrastant les taux de substitution non-synonymes et synonymes (c.f. partie III.2. et III.3).

Ces différents jeux de données permettront de mener des analyses évolutives plus poussées sur le **domaine LRR spécifiquement** comme énoncé précédemment. Plusieurs mécanismes ont déjà pu être identifiés : duplications et délétions de motifs, conversion génique, fusion de domaines, sélection diversifiante (Parniske et al. 1997; Noel et al. 1999). Avec ces données il sera possible de documenter précisément quelles sont leur fréquence relative, combien de motifs ils impliquent, quelle partie du domaine est concernée, s'il existe un lien avec la structure intron/exon du gène et également s'il y a des spécificités évolutives de ce domaine selon les familles.

Ce projet a fait l'objet d'une demande de financement à Agropolis Fondation, mais n'a malheureusement pas été retenu. Sa construction m'a permis d'initier de nouvelles collaborations avec des chercheurs de l'unité PHIM (Thomas Kroj et Stéphane De Mita) et également de prendre contact avec Rod Wing (leader du projet IOMAP 'International Oryza Map Alignment Project') et Andy Jones (leader du projet 'PanOryza BBSRC' financé par la NSF).

I.3. Et chez les blés diploïdes ?

Les blés font partie des espèces majeures pour l'alimentation humaine et animale. Le blé dur est une espèce importante dans le pourtour méditerranéen, et est la matière première pour la fabrication des semoules et des pâtes alimentaires. Ces espèces appartiennent à la tribu des *Triticeae* qui inclut les blés, l'orge, le seigle et leurs espèces apparentées. L'étude des gènes de résistance chez ces espèces et leur utilisation en sélection a commencé bien avant l'avènement du clonage. Aujourd'hui plusieurs d'entre eux ont été clonés et beaucoup appartiennent à la famille des NLR (Sanchez-Martin and Keller 2021). Si les NLR ont été inventoriés depuis chez le blé tendre, incluant la construction d'un pan-'NLRome' (Steuernagel et al. 2020), les LRR-RLK et les LRR-RLP sont, quant à eux, bien moins décrits (Sanchez-Martin and Keller 2021).

Au sein de mon équipe, le complexe d'espèce du blé dur (tétraploïde) est étudié sur plusieurs axes, notamment l'évolution de la diversité au sein des différents compartiments jalonnant sa domestication (Haudry et al. 2007), et l'exploitation de cette diversité ('pré-breeding', (David et al. 2014)), ainsi que sur les interactions plantes-plantes (Freville et al. 2019; Montazeaud et al. 2022). Les interactions plantes-plantes peuvent servir de médiateur lors des réponses immunitaires des plantes (Pelissier et al. 2021). Outre l'intérêt des récepteurs à LRR dans les réponses liées aux stress, en particulier biotiques, une motivation importante pour travailler sur ces récepteurs concerne la possibilité qu'ils puissent être impliqués, au niveau moléculaire, dans les relations plantes-plantes. Les génomes des blés tétraploïdes proviennent d'évènements d'hybridation entre des espèces ancestrales diploïdes dont il existe des représentants actuels (**Fig. 2**). Grâce à l'évolution des technologies de séquençage, de plus en plus de génomes de ces espèces, particulièrement volumineux, sont assemblés *de novo* et publiés, notamment des espèces diploïdes comme des *Aegilops* (Avni et al. 2022; Li et al. 2022) mais aussi les espèces progénitrices des génomes A (Ling et al. 2018) et D (Luo et al. 2017) des blés actuels (**Fig. 2**). Les séquences de deux génomes tétraploïdes sont également disponibles : l'espèce sauvage progénitrice du blé dur (*T. turgidum* ssp. *dicoccoides*) (Avni et al. 2017) et le blé dur lui-même (*T. turgidum* ssp. *durum*) (Maccaferri et al. 2019).

A plus long terme, j'envisage donc d'entreprendre l'étude des récepteurs à LRR dans ce complexe d'espèces, en commençant par les génomes diploïdes, puis, à terme, en transférant les connaissances acquises sur les génotypes tétraploïdes. Les objectifs à long terme seraient de pouvoir (i) **fournir une annotation exhaustive et qualitative des récepteurs à LRR dans ce complexe d'espèces** (incluant le concept de gène non-canonique) qui sera utile, notamment, dans les études de GWAS ou de QTL (des régions d'intérêt pourront être choisies pour une première expertise) (ii) d'aborder la question de **l'impact de la domestication sur l'évolution de ces gènes** et (iii) à terme, de poser la question de **l'impact de la polyploïdie** pour cette famille déjà très complexe dans les génomes 'simples'.

Un volet méthodologique important devra être considéré pour développer notamment l'outil de transfert d'annotation expertisé entre espèces plus éloignées, comme le riz et le blé. Cette partie pourra faire l'objet de collaboration avec mes collègues généticiens impliqués dans les programmes actuels sur le blé au sein de l'équipe (Hélène Fréville, Jacques David et Muriel Tavaud-Pirra), ainsi qu'avec des partenaires travaillant sur les résistances chez le blé (UMR PHIM et/ou GDEC au-delà).

II. Epissage alternatif et adaptation

Comme je l'ai détaillé dans la partie 'Synthèse des travaux', l'épissage alternatif est un mécanisme moléculaire dont le rôle est assez mal compris. J'ai récemment abordé la question de son rôle potentiel dans l'adaptation via l'impact de la domestication sur sa diversité. Les résultats de ce travail, combinés à l'émergence d'une littérature de plus en plus abondante, convergent vers l'idée que l'épissage alternatif joue un rôle dans l'adaptation (Verta and Jacobs 2022). Je propose de développer mon projet pour **étudier l'épissage alternatif chez les plantes et amener des éléments d'analyse de compréhension de son rôle dans l'adaptation**. Le profil d'épissage peut être défini comme la composition en isoformes d'un transcriptome, c'est-à-dire l'inventaire des isoformes et leurs quantités relatives. À l'heure actuelle, plusieurs questions restent posées.

La première concerne la **variabilité des profils d'épissage entre individus au sein d'une même espèce**. Si on assimile un profil d'épissage à un trait phénotypique (ou un ensemble de traits), on parle ici alors de variabilité phénotypique. Chez l'homme, cette question de la variabilité de l'épissage alternatif entre individus a été abordée il y a déjà quelques années (Hull et al. 2007; Kwan et al. 2008; Lu et al. 2012) et s'avère aujourd'hui au cœur des recherches sur de nombreuses maladies, notamment les cancers (Park et al. 2018). Des études d'association ont permis de mettre en évidence des facteurs génétiques contrôlant la variabilité de l'épissage alternatif de certains gènes, aussi bien en *cis* (Hull et al. 2007) qu'en *trans* (Zhang et al. 2009). Le profil d'épissage alternatif de certains gènes semble aussi héritable (Nembaware et al. 2008), ainsi que des évènements de rétention d'intron, potentiellement reliés à l'efficacité du processus de dégradation des ARNm non-sens (Seoighe and Gehring 2010). Chez les plantes, il y a moins de données sur la variabilité intraspécifique de l'épissage alternatif. Une étude chez la vigne a exploré le patron d'épissage alternatif de 10 cultivars (Potenza et al. 2015). Les différentes isoformes se sont avérées plutôt conservées entre les individus et 21% d'entre elles sont présentes dans les 10 individus, malgré le fait que dans la plupart des cas (70% environ), pour un gène donné, une isoforme est exprimée au moins 10 fois moins que la forme canonique. Une étude récente réalisée sur plusieurs centaines d'écotypes d'*Arabidopsis tahliana* a permis, quant à elle, d'identifier des SNP liés aux différences de profils d'épissage (Khokhar et al. 2019). Chez les plantes, il reste à approfondir cette question de l'héritabilité de l'épissage alternatif, et à déterminer si les variations observées sont dues à des variations des sites d'épissage eux-mêmes, ou à proximité, ou à des facteurs de régulations ayant des effets pléiotropes, ou aux deux. Si les variations intraspécifiques des profils d'épissage existent et ont une composante génétique alors elles peuvent être des cibles de la sélection.

La deuxième question est celle du rôle de l'épissage alternatif dans la plasticité phénotypique. L'épissage alternatif est globalement fortement augmenté chez les plantes soumises à un stress environnemental biotique ou abiotique (Staiger and Brown 2013; Filichkin et al. 2015). Ces changements sont quantitatifs, mais aussi qualitatifs, c'est-à-dire que des isoformes différentes peuvent apparaitre (Martin et al. 2021). Ces résultats suggèrent que l'épissage alternatif est une **composante de la réponse plastique des plantes aux stress**. Néanmoins, comment cette réponse se met en place (quels sont les gènes affectés), est-ce qu'elle est variable entre individus (interactions génotype environnement) et quels sont les facteurs génétiques (régulateurs), s'ils existent, qui la contrôlent, restent des questions ouvertes.

La troisième question qui se pose est la part de ces évènements d'épissage alternatif qui ont un **rôle fonctionnel avéré** et quelle est la part de ces évènements qui résulte d'erreurs d'épissages. Un moyen d'explorer cette question est d'étudier le niveau de conservation entre espèces (Barbosa-Morais et al. 2012). Les études sur le niveau de conservation des évènements d'épissage alternatif sont encore peu nombreuses sur les plantes et aboutissent à des résultats parfois contradictoires (Zhang et al. 2015). De plus, ces études sont souvent faites sur des espèces phylogénétiquement assez éloignées, la plupart du temps largement au-delà du genre, ce qui réduit la portée des analyses car les comparaisons d'épissage alternatif sont restreintes aux gènes pour lesquels une relation d'orthologie claire a pu être établie. Ces comparaisons souffrent également souvent de l'hétérogénéité des données (effort de séquençage trop faible, organes échantillonnés différents, stades de développement et environnements considérés hétérogènes, outils d'analyse peu adaptés pour établir les profils d'épissage alternatif) (Severing et al. 2009; Zhang et al. 2015). Enfin, les questions de **l'origine des nouvelles isoformes** et de la vitesse d'évolution (taux d'apparition ou de disparition) sont encore peu documentées. Les relations entre l'épissage alternatif et des paramètres évolutifs comme les taux de substitution, ou les pressions sélectives auxquelles sont soumis les gènes présentant différents profils d'épissage alternatif, ont été peu étudiées chez les plantes.

Pour aborder ces questions, je propose de travailler au sein de l'espèce *Medicago truncatula* pour laquelle les ressources génétiques sont bien caractérisées (Ronfort et al. 2006), et les données génomiques de qualité (Pecrix et al. 2018). Pour aborder plus spécifiquement la troisième question je propose d'étendre les analyses au complexe d'espèce des légumineuses avec deux niveaux de profondeur phylogénétiques, d'abord au sein du genre *Medicago*, puis au sein de plusieurs espèces de légumineuses cultivées (pois, trèfle, lotier, soja, pois chiche …).

II.1. Etude des composantes génétiques et environnementales de l'épissage alternatif.

Les objectifs de cette partie seront de documenter le niveau et les composantes de la diversité de l'épissage alternatif chez *Medicago truncatula* et de fournir une première mesure de son héritabilité.

Dans un premier temps, je propose de construire un dispositif expérimental permettant d'accéder au profil d'épissage de plusieurs génotypes obtenus dans les mêmes conditions, c'est-à-dire à partir des mêmes organes prélevés sur des plantes mises en culture dans un environnement identique, homogène et optimal (non stressant). Une façon simple de procéder sera de travailler avec des graines germées. L'idée est de s'appuyer sur les core-collections de *Medicago truncatula* pour choisir un panel d'individus représentatifs de la variabilité naturelle au sein de l'espèce. Plusieurs core-collections emboitées, de 8, 16, 48 ou 92 individus, sont disponibles (Ronfort et al. 2006). De plus, *Medicago truncatula* étant une espèce autogame, les génotypes sont des lignées fixées. En utilisant plusieurs graines représentant la même lignée génétique, il sera possible de faire des répétitions intra génotype.

D'un point de vue technique, la technologie proposée pour séquencer les transcriptomes extraits de ces échantillons est de type PacBio. En effet cette technologie permet d'accéder à des séquences beaucoup plus longues que les technologies de type NGS classiques, et donc d'obtenir la séquence des ARNm de quasi pleine longueur facilitant l'identification des isoformes. Un point important sera de choisir quels seront les **descripteurs des profils d'épissage alternatifs** qui permettront au mieux de rendre compte de la variabilité observée. En effet, plusieurs descripteurs différents pourront être considérés, comme par exemple le nombre d'isoformes par gène, les niveaux d'expression par gène, relatives (ratio) ou absolues (quantitative), mais aussi les types d'évènements, *i.e.* s'agit-il d'isoformes issues d'évènement de type rétention d'intron ou exclusion d'exons ou autre (**Fig. 14**). Il est tout à fait possible par exemple, que pour certains gènes le nombre d'isoformes soit le même, mais que les ratios d'expression varient, comme nous l'avons montré chez le sorgo [**P24**]. Des développements méthodologiques, en particulier bio-informatiques, seront probablement nécessaires.

Le processus d'épissage est décrit comme comprenant une part de bruit stochastique lié au fonctionnement de la machinerie d'épissage elle-même. Déterminer **quelle est la part de ce bruit et quelle est la part contrôlée génétiquement reste un challenge**. En particulier chez les plantes très peu de données sont publiées sur le niveau de répétabilité des profils d'épissage alternatifs, des répétitions n'étant que peu incluses dans les plans expérimentaux, même lorsque plusieurs conditions sont examinées. Si les résultats confirment l'existence d'une variabilité entre génotypes dans le processus d'épissage alternatif, au moins pour certains de ses descripteurs, il sera alors opportun de déterminer quelle est la **part de la variabilité de ces profils d'épissage qui est d'origine génétique**. Le dispositif proposé incluant des répétitions intra génotype, nous pourrons estimer l'héritabilité, c'est-à-dire la part de cette variation qui est contrôlée génétiquement, en contrastant la variabilité observée entre les génotypes (inter génotype) et la variabilité totale. Une fois de plus, ces analyses devront être faites sur plusieurs composantes de l'épissage alternatif. Il est possible que la machinerie d'épissage produise plus de bruit sur la rétention d'introns que sur l'exclusion d'exons par exemple.

Dans un troisième temps, le cas échéant, je propose d'étudier quels sont les **déterminants génétiques** contrôlant cette diversité par des approches de génétique d'association notamment, pour identifier des régions régulatrices en cis- ou en trans-. Chez Arabidopsis, les différences de profils d'épissage alternatif semblent être davantage cis-régulées par des variations au niveau même des sites d'épissage (Wang et al. 2019). D'autres études ont mis en évidence le rôle de facteurs de régulation agissant plutôt en trans- notamment en impliquant des protéines de type RNA-binding protéines ou des facteurs d'épissage (cités dans (Verta and Jacobs 2022)). La question de l'existence de régions qui agiraient en trans- est particulièrement intéressante car elle suppose une action qui pourrait être de type pléiotropique sur l'épissage alternatif.

Enfin, pour compléter ce volet intraspécifique, il serait intéressant de mener l'ensemble des expérimentations détaillées ci-dessus en utilisant, cette fois, des environnements différents, en particulier des **environnements dans lesquels on applique un stress**. Les profils d'épissage alternatifs pourraient être produits à partir de plantes ayant subi un stress abiotique de type carence hydrique, azotée ou stress salin, pour étudier comment l'épissage alternatif se met en place en réponse à un stress. Ce type de dispositif permettrait notamment de documenter comment l'épissage alternatif intervient dans la réponse plastique des plantes à leur environnement et aussi de déterminer si des interactions entre génotype et environnement (GxE) existent pour certaines de ses composantes.

Ces études seraient menées notamment en collaborations avec les membres de l'équipe ayant des compétences en génétique des populations (GWAS) et en génétique quantitative (héritabilité et GxE), (notamment Laurène Gay et Muriel Tavaud).

II.2. Conservation de l'épissage alternatif à l'échelle interspécifique

Dans cette partie je propose d'utiliser **une démarche comparative pour étudier l'évolution des profils d'épissage alternatif à différentes échelles phylogénétiques** et d'accumuler des indices permettant d'identifier des isoformes candidates pour des investigations du rôle biologique de l'épissage alternatif.

Documenter les niveaux de conservation de l'épissage alternatif au niveau interspécifique est un moyen d'identifier des évènements très conservés pouvant être le signe d'une pression sélective s'étant exercée sur ces évènements au cours de l'évolution et révéler ceux dont la fonction pourrait être importante pour le fonctionnement de la plante. Dans cette partie je propose d'étudier le niveau de conservation des évènements d'épissage alternatif sur un échantillon d'espèces centré autour de l'espèce *Medicago truncatula*, et représentants différents 'pas de temps' évolutifs (clade phylogénétique, genre, famille) chez les légumineuses (**encadré 1**).

Un projet 'starter' financé par le département BAP (projet 'CAS' pour 'Conservation of Alternative Splicing') m'a permis d'initier ces travaux. La première étape a consisté à mettre au point la production des données moléculaires. Nous avons décidé d'axer toute notre démarche technique sur l'utilisation de la technologie de séquençage PacBio (collègues de l'INRAE de Clermont-Ferrand) combinée à la technologie TeloPrime (Lexogen). Ces mises au point ont été réalisées en collaboration avec Sylvain Santoni (ingénieur de recherche dans l'équipe, responsable du plateau de biologie moléculaire), Audrey Weber et Muriel Latreille (techniciennes au laboratoire de biologie moléculaire). Les premiers jeux de données ont été produits via la technologie Isoseq et nous sommes en train de développer des scripts d'analyse spécifiques (collaboration avec Vincent Ranwez et Gautier Sarah, UMR AGAP).

Parallèlement, nous avons contacté plusieurs laboratoires pour récupérer les ressources génétiques : l'USDA via le projet HapMap pour le genre *Medicago* (contact Nevin Young), l'Université de Purdue pour le soja, l'Université de Tokyo pour le lotier, nos collègues de l'INRAE de Dijon pour le pois, l'ICRISAT pour le pois chiche, Agresearch en Nouvelle-Zélande pour le trèfle. Ces différentes démarches nous ont permis de récupérer des lots de grains de ces différentes espèces.

L'analyse des résultats pourra se faire dans un cadre évolutif clairement identifié nous permettant ainsi de tracer **l'histoire évolutive des évènements d'épissage alternatif** (ancestraux ou récents, taux de perte/acquisition, occurrences multiples ou/et évolution convergente) et de mettre en relation ces évènements liés à l'épissage alternatif avec d'autres paramètres (taux de mutations, contraintes sélectives).

Plus précisément, je tâcherais de répondre aux questions suivantes : **est-ce que le niveau de conservation est équivalent pour les différents types d'évènements d'épissage alternatif** ? Est-ce que les évènements qui ne modifient pas le cadre de lecture sont significativement plus conservés ? Quels sont les **taux d'apparition/perte des différents types d'évènements** ? Quelles sont les caractéristiques clés (si elles existent) d'un évènement d'épissage alternatif préservé et peut-on prédire au moins partiellement la **transférabilité** d'un évènement d'épissage alternatif d'une espèce (modèle) à une autre (d'intérêt). La question de la transférabilité des connaissances sur l'épissage alternatif d'une espèce vers une autre reste cruciale pour se focaliser sur les études les plus prometteuses sans explorer de trop nombreuses conditions environnementales.

Enfin, un point important qui sera abordé dans les deux parties de ce projet (intra- et interspécifique) sera de déterminer quelles sont les isoformes à partir desquelles des protéines potentielles peuvent être traduites et quelles sont celles qui contiennent des ruptures de phase ou des codons stop, c'est-à-dire qui correspondent à des ARN non-sens. Une perspective à plus long terme de ce travail pourrait être de se rapprocher de collectifs spécialisés dans les analyses fonctionnelles, soit via des banques de mutants, soit par analyse protéique, pour aller jusqu'à l'identification des fonctions potentielles des isoformes candidates identifiées.

Postface

Pour conclure ce document, je voudrais aborder quelques points de réflexion. Le premier est directement lié aux deux axes de mon projet. Les duplications de gènes et l'épissage alternatif sont deux mécanismes qui permettent de générer de la complexité au sein des génomes. En augmentant le nombre de copies d'un gène, la duplication crée un état de redondance sur lequel des forces évolutives différentes vont pouvoir s'exercer. Nous avons montré que cet état de redondance est souvent le support pour l'adaptation et cela est particulièrement fréquent dans la famille des LRR-RLK. Sans modifier le nombre de copies, l'épissage alternatif, quant à lui, complexifie la relation entre génotype et phénotype. Ce mécanisme n'a pas encore bénéficié d'autant de travaux théoriques et empiriques mais il est de plus en plus étudié. S'il a été montré qu'une corrélation négative existe entre la propension d'un gène à exprimer plusieurs isoformes et sa propension à se dupliquer, une question légitime qui émerge de mon projet sera celle de l'occurrence de l'épissage alternatif dans les récepteurs à LRR. En particulier, dans certains gènes, le domaine LRR est organisé de façon intrigante : la partie codante contient autant d'exons que de motifs LRR. Une manière très 'simple' de modifier le domaine LRR serait alors d'épisser de façon alternative ces exons là pour créer des agencements différents de LRR.

Le deuxième point que je voudrais aborder est lié à la façon exponentielle dont les données de séquence sont produites aujourd'hui, principalement depuis l'avènement des dites 'nouvelles générations de séquençages', qui ne cessent d'évoluer. Face à cette augmentation quantitative et qualitative des données de séquence, les outils d'analyses (annotation, comparaisons, assemblage, génomique des populations, etc.) doivent s'adapter constamment, se développer et être de plus en plus performants. L'enchaînement de ces outils, nécessaires pour passer de la donnée aux résultats interprétables, permet de calibrer et de répéter l'ensemble des analyses, assurant une forte traçabilité, en théorie (conservation des données, 'versionning' des scripts et outils). Les points de contrôle qui jalonnent ces étapes sont très importants pour détecter d'éventuelles anomalies, ou biais, à l'origine de résultats qui pourraient être interprétés de façon erronée. Ces contrôles se passent de plus en plus de l'expertise directe de l'œil humain. Il me semble néanmoins important de garder des 'fenêtres de contrôle' à échelle humaine sur les données (par échantillonnage aléatoire par exemple, ou en sélectionnant et examinant les données atypiques). En effet, l'observation, notamment sans a priori, permet souvent de voir des 'choses inattendues' qui peuvent poser problème ou susciter l'intérêt, mais qui ne sont pas détectées par les outils tant qu'ils ne sont pas programmés pour le faire.

Malgré une grande hétérogénéité de qualité dans les données de séquences publiques, certains auteurs privilégient la qualité et l'accessibilité. Dans ce contexte, il est possible aujourd'hui d'envisager des parties entières de projets qui ne nécessitent plus de production de données, mais qui proposent d'exploiter celles qui sont mises à la disposition de la communauté scientifique. Cette démarche a deux avantages qui me semblent non négligeables. Le premier est qu'il permet de réduire l'impact environnemental des projets. Le deuxième est qu'il permet d'allouer les ressources ainsi économisées pour financer des salaires et des indemnités de stage. Une plus-value incontestable sera apportée par l'augmentation du niveau d'expertise des analyses qui seront faites sur ces données déjà existantes.

Pour finir, je reviendrai juste sur l'importance des interactions entre les acteurs de la recherche, en particulier entre les différents métiers, qui sont tous indispensables pour que les programmes aboutissent et pour que de nouvelles idées émergent et se concrétisent. Parmi ces acteurs, les étudiants sont une force incontestable. Au-delà du fait qu'ils permettent concrètement de faire avancer les projets et qu'il est souvent très satisfaisant de transmettre méthodes et connaissances, ils nous permettent indéniablement de nous questionner différemment et ainsi d'entretenir et de perpétrer le processus de réflexion qui nous permet de progresser.

RÉFÉRENCES CITÉES

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**: 135-141.

Akoz G, Nordborg M. 2019. The Aquilegia genome reveals a hybrid origin of core eudicots. *Genome Biol* **20**: 256.

Alix K, Gerard PR, Schwarzacher T, Heslop-Harrison JSP. 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann Bot* **120**: 183-194.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Avni R, Lux T, Minz-Dub A, Millet E, Sela H, Distelfeld A, Deek J, Yu G, Steuernagel B, Pozniak C et al. 2022. Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. *Plant J* doi:10.1111/tpj.15664.

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K et al. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**: 93-97.

Azani N, Babineau M, Bailey CD, Banks H, Barbosa AR, Pinto RB, Boatwright JS, Borges LM, Brown GK, Bruneau A et al. 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* **66**: 44-77.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587-1593.

Barragan AC, Weigel D. 2021. Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell* **33**: 814-831.

Bartok O, Kyriacou CP, Levine J, Sehgal A, Kadener S. 2013. Adaptation of molecular circadian clockwork to environmental changes: a role for alternative splicing and miRNAs. *Proc Biol Sci* **280**: 20130011.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nat Plants* **6**: 914-920.

Bebber DP, Ramotowski MAT, Gurr SJ. 2013. Crop pests and pathogens move polewards in a warming world. *NATURE CLIMATE CHANGE* **3**: 985-988.

Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, Minio A, Rau D, Rodriguez M, Panziera A et al. 2014. Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. *Plant Cell* **26**: 1901-1912.

Blencowe BJ. 2017. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem Sci* **42**: 407-408.

Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J. 2009. The 'inner circle' of the cereal genomes. *Curr Opin Plant Biol* **12**: 119-125.

Caceres JF, Kornblihtt AR. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* **18**: 186-193.

Casola C, Betran E. 2017. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? *Genome Biol Evol* **9**: 1351-1373.

Chaudhary S, Khokhar W, Jabre I, Reddy ASN, Byrne LJ, Wilson CM, Syed NH. 2019. Alternative Splicing and Protein Diversity: Plants Versus Animals. *Front Plant Sci* **10**: 708.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.

Chen T. 2021. Identification and characterization of the LRR repeats in plant LRR-RLKs. *BMC Mol Cell Biol* **22**: 9.

Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* **4**: 258-268.

Cieply B, Carstens RP. 2015. Functional roles of alternative splicing factors in human disease. *Wiley interdisciplinary reviews RNA* **6**: 311-326.

D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC. 1996. Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Mol Gen Genet* **250**: 405-413.

David J, Holtz Y, Ranwez V, Santoni S, Sarah G, Ardisson M, Poux G, Choulet F, Genthon C, Roumet P et al. 2014. Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. *Molecular Breeding* **34**: 1531-1548.

De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki R, Bataillon T. 2007. Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in Medicago truncatula. *Genetics* **177**: 2123-2133.

De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet* **13**: 27.

Defoort J, Van de Peer Y, Carretero-Paulet L. 2019. The Evolution of Gene Duplicates in Angiosperms and the Impact of Protein-Protein Interactions and the Mechanism of Duplication. *Genome Biol Evol* **11**: 2292-2305.

Demuth JP, Hahn MW. 2009. The life and death of gene families. *Bioessays* **31**: 29-39.

Dievart A, Clark SE. 2003. Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Curr Opin Plant Biol* **6**: 507-516.

Dievart A, Clark SE. 2004. LRR-containing receptors regulating plant development and defense. *Development* **131**: 251-261.

Dievart A, Perin C, Hirsch J, Bettembourg M, Lanau N, Artus F, Bureau C, Noel N, Droc G, Peyramard M et al. 2016. The phenome analysis of mutant alleles in Leucine-Rich Repeat Receptor-Like Kinase genes in rice reveals new potential targets for stress tolerant cereals. *Plant Sci* **242**: 240-249.

Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. 2014. Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics* **15**: 431.

Dolatabadian A, Patel DA, Edwards D, Batley J. 2017. Copy number variation and disease resistance in plants. *Theor Appl Genet* **130**: 2479-2490.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. *Mol Biol Evol* **29**: 1861-1874.

Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML. 2012. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol* **29**: 2265-2283.

Feldman M, Levy AA, Fahima T, Korol A. 2012. Genomic asymmetry in allopolyploid plants: wheat as a model. *J Exp Bot* **63**: 5045-5059.

Feng J, Li J, Gao Z, Lu Y, Yu J, Zheng Q, Yan S, Zhang W, He H, Ma L et al. 2015. SKIP Confers Osmotic Tolerance during Salt Stress by Controlling Alternative Gene Splicing in Arabidopsis. *Mol Plant* **8**: 1038-1052.

Feuillet C, Keller B. 2002. Comparative genomics in the grass family: molecular characterization of Grass genome structure and evolution. *Annals of Botany* **89**: 3-10.

Filichkin S, Priest HD, Megraw M, Mockler TC. 2015. Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. *Curr Opin Plant Biol* **24**: 125-135.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99-113.

Fitch WM. 2000. Homology a personal view on some of the problems. *Trends Genet* **16**: 227-231.

Freville H, Roumet P, Rode NO, Rocher A, Latreille M, Muller MH, David J. 2019. Preferential helping to relatives: A potential mechanism responsible for lower yield of crop variety mixtures? *Evol Appl* **12**: 1837-1849.

Gabaldon T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14**: 360-366.

Gale MD, Devos KM. 1998. Comparative genetics in the grasses. *Proc Natl Acad Sci USA* **95**: 1971-1974.

Gautier MF, Cosson P, Guirao A, Alary R, Joudrier P. 2000. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid *Triticum* species. *Plant Science* **153**: 81-91.

Glemin S, Bataillon T. 2009. A comparative view of the evolution of grasses under domestication. *New Phytol* **183**: 273-290.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**: 725-736.

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC. 1996. RFLP mapping in cultivated sugarcane (Saccharum spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics* **142**: 987-1000.

Gros-Balthazard M, Besnard G, Sarah G, Holtz Y, Leclercq J, Santoni S, Wegmann D, Glemin S, Khadari B. 2019. Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *Plant J* **100**: 143-157.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307-321.

Guyon-Debast A, Alboresi A, Terret Z, Charlot F, Berthier F, Vendrell-Mir P, Casacuberta JM, Veillet F, Morosinotto T, Gallois JL et al. 2021. A blueprint for gene function analysis through Base Editing in the model plant Physcomitrium (Physcomitrella) patens. *New Phytol* **230**: 1258-1272.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* **148**: 993-1003.

Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glemin S et al. 2007. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol Biol Evol* **24**: 1506-1517.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**: D309-D314.

Hull J, Campino S, Rowlands K, Chan MS, Copley RR, Taylor MS, Rockett K, Elvidge G, Keating B, Knight J et al. 2007. Identification of common genetic variation that modulates alternative splicing. *PLoS Genet* **3**: e99.

Ilic K, SanMiguel PJ, Bennetzen JL. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci USA* **100**: 12265-12270.

Iniguez LP, Hernandez G. 2017. The Evolutionary Relationship between Alternative Splicing and Gene Duplication. *Front Genet* **8**: 14.

Innan H. 2009. Population genetic models of duplicated genes. *Genetica* **137**: 19-37.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**: 97-108.

International Brachypodium I. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**: 763-768.

Irimia M, Roy SW. 2014. Origin of Spliceosomal Introns and Alternative Splicing. *Cold Spring Harbor Perspectives in Biology* **6**.

Jain R, Jenkins J, Shu S, Chern M, Martin JA, Copetti D, Duong PQ, Pham NT, Kudrna DA, Talag J et al. 2019. Genome sequence of the model rice variety KitaakeX. *BMC Genomics* **20**: 905.

Jiang WK, Liu YL, Xia EH, Gao LZ. 2013. Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. *Plant Physiol* **161**: 1844-1861.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97-100.

Kazan K. 2003. Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. *Trends Plant Sci* **8**: 468-471.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol Evol* **24**: 572-582.

Khokhar W, Hassan MA, Reddy ASN, Chaudhary S, Jabre I, Byrne LJ, Syed NH. 2019. Genome-Wide Identification of Splicing Quantitative Trait Loci (sQTLs) in Diverse Ecotypes of Arabidopsis thaliana. *Front Plant Sci* **10**: 1160.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* **44**: D73-80.

Koonin EV. 2009. Evolution of genome architecture. *Int J Biochem Cell Biol* **41**: 298-306.

Kopelman NM, Lancet D, Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37**: 588-589.

Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**: 153-165.

Kourelis J, van der Hoorn RAL. 2018. Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. *Plant Cell* **30**: 285-299.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**: 225-231.

Lehti-Shiu MD, Shiu SH. 2012. Diversity, classification and function of the plant protein kinase superfamily. *Philos Trans R Soc Lond B Biol Sci* **367**: 2619-2639.

Lehti-Shiu MD, Zou C, Hanada K, Shiu SH. 2009. Evolutionary History and Stress Regulation of Plant Receptor-Like Kinase/Pelle Genes. *Plant Physiology* **150**: 12-26.

Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol* **194**: 629-646.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**: 189-192.

Li LF, Zhang ZB, Wang ZH, Li N, Sha Y, Wang XF, Ding N, Li Y, Zhao J, Wu Y et al. 2022. Genome sequences of five Sitopsis species of Aegilops and the origin of polyploid wheat B subgenome. *Mol Plant* **15**: 488-503.

Lijavetzky D, Muzzi G, Wicker T, Keller B, Wing R, Dubcovsky J. 1999. Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* **42**: 1176-1182.

Linard B, Ebersberger I, McGlynn SE, Glover N, Mochizuki T, Patricio M, Lecompte O, Nevers Y, Thomas PD, Gabaldon T et al. 2021. Ten Years of Collaborative Progress in the Quest for Orthologs. *Mol Biol Evol* **38**: 3033-3045.

Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y et al. 2018. Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature* **557**: 424-428.

Liu W, Chen L, Zhang S, Hu F, Wang Z, Lyu J, Wang B, Xiang H, Zhao R, Tian Z et al. 2019. Decrease of gene expression diversity during domestication of animals and plants. *BMC Evol Biol* **19**: 19.

Lopez-Diez R, Rastrojo A, Villate O, Aguado B. 2013. Complex tissue-specific patterns and distribution of multiple RAGE splice variants in different mammals. *Genome Biol Evol* **5**: 2420-2435.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* **102**: 10557-10562.

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W et al. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* **20**: 1238-1249.

Lu ZX, Jiang P, Xing Y. 2012. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley interdisciplinary reviews RNA* **3**: 581-592.

Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, Huo N, Zhu T, Wang L, Wang Y et al. 2017. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* **551**: 498-502.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG et al. 2019. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* **51**: 885-895.

Mandakova T, Lysak MA. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Curr Opin Plant Biol* **42**: 55-65.

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res* **22**: 1184-1195.

Martin G, Marquez Y, Mantica F, Duque P, Irimia M. 2021. Alternative splicing landscapes in Arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals. *Genome Biol* **22**: 35.

Martinez O, Arce-Rodriguez ML, Hernandez-Godinez F, Escoto-Sandoval C, Cervantes-Hernandez F, Hayano-Kanashiro C, Ordaz-Ortiz JJ, Reyes-Valdes MH, Razo-Mendivil FG, Garces-Claver A et al. 2021. Transcriptome Analyses Throughout Chili Pepper Fruit Development Reveal Novel Insights into the Domestication Process. *Plants (Basel)* **10**.

McGlincy NJ, Smith CWJ. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences* **33**: 385-393.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet* **14**: 840-852.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**: 541.

Michael TP, VanBuren R. 2020. Building near-complete plant genomes. *Curr Opin Plant Biol* **54**: 26-33.

Michelmore RW, Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* **8**: 1113-1130.

Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH et al. 1998. Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* **150**: 1663-1682.

Mizuno H, Katagiri S, Kanamori H, Mukai Y, Sasaki T, Matsumoto T, Wu J. 2020. Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. *Sci Rep* **10**: 872.

Modrek B, Lee C. 2002. A genomic view of alternative splicing. *Nat Genet* **30**: 13-19.

Montazeaud G, Flutre T, Ballini E, Morel JB, David J, Girodolle J, Rocher A, Ducasse A, Violle C, Fort F et al. 2022. From cultivar mixtures to allelic mixtures: opposite effects of allelic richness between genotypes and genotype richness in wheat. *New Phytol* **233**: 2573-2584.

Monteiro F, Nishimura MT. 2018. Structural, Functional, and Genomic Diversity of Plant NLR Proteins: An Evolved Resource for Rational Engineering of Plant Immunity. *Annu Rev Phytopathol* **56**: 243-267.

Moullet O, Zhang HB, Lagudah ES. 1999. Construction and characterisation of a large DNA insert library from the D genome of wheat. *Thoer Appl Genet* **99**: 305-313.

Muller MH, Poncet C, Prosperi JM, Santoni S, Ronfort J. 2006. Domestication history in the Medicago sativa species complex: inferences from nuclear sequence polymorphism. *Mol Ecol* **15**: 1589-1602.

Murat F, Armero A, Pont C, Klopp C, Salse J. 2017. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* **49**: 490-496.

Mussurova S, Al-Bader N, Zuccolo A, Wing RA. 2020. Potential of Platinum Standard Reference Genomes to Exploit Natural Variation in the Wild Relatives of Rice. *Front Plant Sci* **11**: 579980.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**: 121-152.

Nelson R, Wiesner-Hanks T, Wisser R, Balint-Kurti P. 2018. Navigating complexity to breed disease-resistant crops. *Nat Rev Genet* **19**: 21-33.

Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C. 2008. Genome-wide survey of allele-specific splicing in humans. *BMC Genomics* **9**: 265.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457-463.

Noel L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD. 1999. Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. *Plant Cell* **11**: 2099-2112.

Ohno S. 1970. Evolution by gene duplication. *George Allen and Unwin, London*.

Panahi B, Mohammadi SA, Ebrahimi Khaksefidi R, Fallah Mehrabadi J, Ebrahimie E. 2015. Genome-wide analysis of alternative splicing events in Hordeum vulgare: Highlighting retention of intron-based splicing and its possible function through network analysis. *FEBS Lett* **589**: 3564-3575.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of Gene Duplication in Plants. *Plant Physiol* **171**: 2294-2316.

Parisod C, Badaeva ED. 2020. Chromosome restructuring among hybridizing wild wheats. *New Phytol* **226**: 1263-1273.

Park E, Pan Z, Zhang Z, Lin L, Xing Y. 2018. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**: 11-26.

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BB, Jones JD. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* **91**: 821-832.

Pecrix Y, Staton SE, Sallet E, Lelandais-Briere C, Moreau S, Carrere S, Blein T, Jardinaud MF, Latrasse D, Zouine M et al. 2018. Whole-genome landscape of Medicago truncatula symbiotic genes. *Nat Plants* **4**: 1017-1025.

Pelissier R, Violle C, Morel JB. 2021. Plant immunity: Good fences make good neighbors? *Curr Opin Plant Biol* **62**: 102045.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* **27**: 1759-1767.

Pilet-Nayel ML, Moury B, Caffier V, Montarry J, Kerlan MC, Fournet S, Durel CE, Delourme R. 2017. Quantitative Resistance to Plant Pathogens in Pyramiding Strategies for Durable Crop Protection. *Front Plant Sci* **8**: 1838.

Potenza E, Racchi ML, Sterck L, Coller E, Asquini E, Tosatto SC, Velasco R, Van de Peer Y, Cestaro A. 2015. Exploration of alternative splicing events in ten different grapevine cultivars. *BMC Genomics* **16**: 706.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol* **20**: 38.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* **6**: e22594.

Rapp RA, Haigler CH, Flagel L, Hovav RH, Udall JA, Wendel JF. 2010. Gene expression in developing fibres of Upland cotton (Gossypium hirsutum L.) was massively altered by domestication. *BMC Biol* **8**: 139.

Reixachs-Sole M, Eyras E. 2022. Uncovering the impacts of alternative splicing on the proteome with current omics techniques. *Wiley interdisciplinary reviews RNA* doi:10.1002/wrna.1707: e1707.

Romiguier J, Figuet E, Galtier N, Douzery EJ, Boussau B, Dutheil JY, Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* **7**: e33852.

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi JM. 2006. Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. *BMC Plant Biol* **6**: 28.

Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Perin C, Conte MG. 2011. GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res* **39**: D1095-1102.

Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, Stacey G, Doerge RW, Jackson SA. 2013. The fate of duplicated genes in a polyploid plant genome. *Plant J* **73**: 143-153.

Salse J. 2016. Ancestors of modern plant crops. *Curr Opin Plant Biol* **30**: 134-142.

Sanchez-Martin J, Keller B. 2021. NLR immune receptors and diverse types of non-NLR proteins control race-specific resistance in Triticeae. *Curr Opin Plant Biol* **62**: 102053.

Sauquet H, von Balthazar M, Magallon S, Doyle JA, Endress PK, Bailes EJ, Barroso de Morais E, Bull-Herenu K, Carrive L, Chartier M et al. 2017. The ancestral flower of angiosperms and its early diversification. *Nat Commun* **8**: 16047.

Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. 2019. The global burden of pathogens and pests on major food crops. *Nat Ecol Evol* **3**: 430-439.

Seoighe C, Gehring C. 2010. Heritability in the efficiency of nonsense-mediated mRNA decay in humans. *PLoS One* **5**: e11657.

Severing EI, van Dijk AD, Stiekema WJ, van Ham RC. 2009. Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics* **10**: 154.

Shakhnovich BE, Koonin EV. 2006. Origins and impact of constraints in evolution of gene families. *Genome Res* **16**: 1529-1536.

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**: 1220-1234.

Smith CCR, Tittes S, Mendieta JP, Collier-Zans E, Rowe HC, Rieseberg LH, Kane NC. 2018. Genetics of alternative splicing evolution during sunflower domestication. *Proc Natl Acad Sci U S A* **115**: 6768-6773.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* **35**: 119-125.

Staiger D, Brown JW. 2013. Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* **25**: 3640-3656.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet* **50**: 285-296.

Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek HJ, Yu G, Baggs E, Witek AI, Yadav I, Krasileva KV et al. 2020. The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. *Plant Physiol* **183**: 468-482.

Sun Y, Xiao H. 2015. Identification of alternative splicing events by RNA sequencing in early growth tomato fruits. *BMC Genomics* **16**: 948.

Tack DC, Pitchers WR, Adams KL. 2014. Transcriptome analysis indicates considerable divergence in alternative splicing between duplicated genes in Arabidopsis thaliana. *Genetics* **198**: 1473-1481.

Tang P, Zhang Y, Sun XQ, Tian DC, Yang SH, Ding J. 2010. Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species. *Plant Science* **179**: 399-406.

Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, Van Allen B, Habben J, Li B. 2016. Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. *Plant Physiol* **170**: 586-599.

Thudi M, Palakurthi R, Schnable JC, Chitikineni A, Dreisigacker S, Mace E, Srivastava RK, Satyavathi CT, Odeny D, Tiwari VK et al. 2021. Genomic resources in plant breeding for sustainable agriculture. *J Plant Physiol* **257**: 153351.

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc Natl Acad Sci USA* **96**: 7409-7414.

Tranchant-Dubreuil C, Rouard M, Sabot F. 2019. Plant Pangenome: Impacts On Phenotypes And Evolution. *Annual Plant Reviews, Wiley Online Library 2019* doi:10.1002/9781119312994.apr0664.

Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME. 2021. Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci* **26**: 631-649.

Veillet F, Durand M, Kroj T, Cesari S, Gallois J-L. 2020. Precision Breeding Made Real with CRISPR: Illustration through Genetic Resistance to Pathogens. *Plant Comm* **1**: 100102.

Verta JP, Jacobs A. 2022. The role of alternative splicing in adaptation and evolution. *Trends Ecol Evol* **37**: 299-308.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. *Ann Bot* **120**: 195-207.

Walsh B. 2003. Population-genetic models of the fates of duplicate genes. *Genetica* **118**: 279-294.

Walters B, Lum G, Sablok G, Min XJ. 2013. Genome-wide landscape of alternative splicing events in Brachypodium distachyon. *DNA Res* **20**: 163-171.

Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ. 2015. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. *Sci Rep* **5**: 13957.

Wang J, Tao F, Marowsky NC, Fan C. 2016. Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes. *Plant Physiol* **172**: 427-440.

Wang X, Yang M, Ren D, Terzaghi W, Deng XW, He G. 2019. Cis-regulated alternative splicing divergence and its potential contribution to environmental responses in Arabidopsis. *Plant J* **97**: 555-570.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568-573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.

Young ND Debelle F Oldroyd GE Geurts R Cannon SB Udvardi MK Benedito VA Mayer KF Gouzy J Schoof H et al. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**: 520-524.

Yu K, Feng M, Yang G, Sun L, Qin Z, Cao J, Wen J, Li H, Zhou Y, Chen X et al. 2020. Changes in Alternative Splicing in Response to Domestication and Polyploidization in Wheat. *Plant Physiol* **184**: 1955-1968.

Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV. 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **49**: D389-D393.

Zhang C, Yang H, Yang H. 2015. Evolutionary Character of Alternative Splicing in Plants. *Bioinform Biol Insights* **9**: 47-52.

Zhang J, Chen LL, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song JM, Xie W et al. 2016. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc Natl Acad Sci U S A* **113**: E5163-5171.

Zhang W, Duan S, Bleibel WK, Wisel SA, Huang RS, Wu X, He L, Clark TA, Chen TX, Schweitzer AC et al. 2009. Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum Genet* **125**: 81-93.

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**: 278-284.

Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, Al-Bader N, Sobel-Sorenson C, Parakkal P et al. 2020. A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci Data* **7**: 113.

LISTE DES PRODUCTIONS SCIENTIFIQUES

Articles à comité de lecture

[**P1**] Chantret, N., M. T. Pavoine, and G. Doussinault. **1999**. 'The Race-Specific Resistance Gene to Powdery Mildew, MlRE, Has a Residual Effect on Adult Plant Resistance of Winter Wheat Line RE714', Phytopathology, 89: 533-9.

[**P2**] Chantret, N., P. Sourdille, M. Röder, M. Tavaud, M. Bernard, and G. Doussinault. **2000**. 'Location and mapping of the powdery mildew resistant gene MlRE and detection of a resistance QTL by bulked segregant analysis (BSA) with microsatellites in wheat.', Theor Appl Genet, 100: 1217-24.

[**P3**] Chantret, N., D. Mingeot, P. Sourdille, M. Bernard, J.M. Jacquemin, and G. Doussinault. **2001**. 'A major QTL for powdery mildew resistance is stable over time and at two development stages in winter wheat.', Theor Appl Genet, 103: 962-71.

[**P4**] Mingeot, D., N. Chantret, P. V. Baret, A. Dekeyser, N. Boukhatem, P. Sourdille, G. Doussinault, and J. M. Jacquemin. **2002**. 'Mapping QTL involved in adult plant resistance to powdery mildew in the winter wheat line RE714 in two susceptible genetic backgrounds', Plant Breeding, 121: 133-40.

[**P5**] Cenci, A., N. Chantret, X. Kong, Y. Gu, O. D. Anderson, T. Fahima, A. Distelfeld, and J. Dubcovsky. **2003**. 'Construction and characterization of a half million clone BAC library of durum wheat (Triticum turgidum ssp. durum)', Theor. Appl. Genet., 107: 931-9.

[**P6**] Cenci, A., S. Somma, N. Chantret, J. Dubcovsky, and A. Blanco. **2004**. 'PCR identification of durum wheat BAC clones containing genes coding for carotenoid biosynthesis enzymes and their chromosome localization', Genome, 47: 911-7.

[**P7**] Chantret, N., A. Cenci, F. Sabot, O. Anderson, and J. Dubcovsky. **2004**. 'Sequencing of the Triticum monococcum Hardness locus reveals good microcolinearity with rice', Mol. Genet. Genomics, 271: 377-86.

[**P8**] Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M. F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. **2005**. 'Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (triticum and aegilops)', Plant Cell, 17: 1033-45.

[**P9**] Sabot, F., R. Guyot, T. Wicker, N. Chantret, B. Laubin, B. Chalhoub, P. Leroy, P. Sourdille, and M. Bernard. **2005**. 'Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations', Mol. Genet. Genomics, 274: 119-30.

[**P10**] Sabot, F., P. Sourdille, N. Chantret, and M. Bernard. **2006**. 'Morgane, a new LTR retrotransposon group, and its subfamilies in wheats', Genetica, 128: 439-47.

[**P11**] Jannoo, N., L. Grivet, N. Chantret, O. Garsmeur, J. C. Glaszmann, P. Arruda, and A. D'Hont. **2007**. 'Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome', Plant J, 50: 574-85.

[**P12**] Boutrot, F., N. Chantret, and M. F. Gautier. **2008**. 'Genome-wide analysis of the rice and Arabidopsis non-specific lipid transfer protein (nsLtp) gene families and identification of wheat nsLtp genes by EST data mining', BMC Genomics, 9: 86.

[**P13**] Chantret, N., J. Salse, F. Sabot, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M. F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, P. Leroy, M. Bernard, and B. Chalhoub. (**2008**). 'Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species', J Mol Evol, 66: 138-50.

[**P14**] De Mita, S., N. Chantret, K. Loridon, J. Ronfort, and T. Bataillon. **2011**. 'Molecular adaptation in flowering and symbiotic recognition pathways: insights from patterns of polymorphism in the legume Medicago truncatula', BMC Evol Biol, 11: 229.

[**P15**] Ho-Huu, J., J. Ronfort, S. De Mita, T. Bataillon, I. Hochu, A. Weber, and N. Chantret. **2012**. 'Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.)', BMC Evol Biol, 12: 195.

[**P16**] Loridon, K., C. Burgarella, N. Chantret, F. Martins, J. Gouzy, J. M. Prosperi, and J. Ronfort. **2013**. 'Single-nucleotide polymorphism discovery and diversity in the model legume Medicago truncatula', Mol Ecol Resour, 13: 84-95.

[**P17**] Bonhomme, M., O. Andre, Y. Badis, J. Ronfort, C. Burgarella, N. Chantret, J. M. Prosperi, R. Briskine, J. Mudge, F. Debelle, H. Navier, H. Miteul, A. Hajri, A. Baranger, P. Tiffin, B. Dumas, M. L. Pilet-Nayel, N. D. Young, and C. Jacquet. **2014**. 'High-density genome-wide association mapping implicates an F-box encoding gene in Medicago truncatula resistance to Aphanomyces euteiches', New Phytol, 201: 1328-42.

[**P18**] Fischer, I., J. Dainat, V. Ranwez, S. Glemin, J. F. Dufayard, and N. Chantret. **2014**. 'Impact of recurrent gene duplication on adaptation of plant genomes', BMC Plant Biol, 14: 151.

[**P19**] Burgarella, C., N. Chantret, L. Gay, J. M. Prosperi, M. Bonhomme, P. Tiffin, N. D. Young, and J. Ronfort. **2016**. 'Adaptation to climate through flowering phenology: a case study in Medicago truncatula', Mol Ecol, 25: 3397-415.

[**P20**] Fischer, I., A. Dievart, G. Droc, J. F. Dufayard, and N. Chantret. **2016**. 'Evolutionary Dynamics of the Leucine-Rich Repeat Receptor-Like Kinase (LRR-RLK) Subfamily in Angiosperms', Plant Physiol, 170: 1595-610.

[**P21**] Karaki, L., P. Da Silva, F. Rizk, C. Chouabe, N. Chantret, V. Eyraud, F. Gressent, C. Sivignon, I. Rahioui, D. Kahn, C. Brochier-Armanet, Y. Rahbe, and C. Royer. **2016**. 'Genome-wide analysis identifies gain and loss/change of function within the small multigenic insecticidal Albumin 1 family of Medicago truncatula', BMC Plant Biol, 16: 63.

[**P22**] Sarah, G., F. Homa, S. Pointet, S. Contreras, F. Sabot, B. Nabholz, S. Santoni, L. Saune, M. Ardisson, N. Chantret, C. Sauvage, J. Tregear, C. Jourda, D. Pot, Y. Vigouroux, H. Chair, N. Scarcelli, C. Billot, N. Yahiaoui, R. Bacilieri, B. Khadari, M. Boccara, A. Barnaud, J. P. Peros, J. P. Labouisse, J. L. Pham, J. David, S. Glemin, and M. Ruiz. **2016**. 'A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives', Mol Ecol Resour.

[**P23**] Dufayard, J. F., M. Bettembourg, I. Fischer, G. Droc, E. Guiderdoni, C. Perin, N. Chantret, and A. Dievart. **2017**. 'New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms', Front Plant Sci, 8: 381.

[**P24**] Ranwez, V., A. Serra, D. Pot, and N. Chantret. **2017**. 'Domestication reduces alternative splicing expression variations in sorghum', PLoS ONE, 12: e0183454.

[**P25**] Cenci, A., N. Chantret, and M. Rouard. **2018**. 'Glycosyltransferase Family 61 in Liliopsida (Monocot): The Story of a Gene Family Expansion', Front Plant Sci, 9: 1843.

[**P26**] Plomion, C., J.-M. Aury, J. Amselem, T. Leroy, F. Murat, S. Duplessis, S. Faye, N. Francillonne, K. Labadie, G. Le Provost, I. Lesur, J. Bartholomé, P. Faivre-Rampant, A. Kohler, J.-C. Leplé, N. Chantret, J. Chen, A. Diévart, T. Alaeitabar, V. Barbe, C. Belser, H. Bergès, C. Bodénès, M.-B. Bogeat-Triboulot, M.-L. Bouffaud, B. Brachi, E. Chancerel, D. Cohen, A. Couloux, C. Da Silva, C. Dossat, F. Ehrenmann, C. Gaspin, J. Grima-Pettenati, E. Guichoux, A. Hecker, S. Herrmann, P. Hugueney, I. Hummel, C. Klopp, C. Lalanne, M. Lascoux, E. Lasserre, A. Lemainque, M.-L. Desprez-Loustau, I. Luyten, M.-A. Madoui, S. Mangenot, C. Marchal, F. Maumus, J. Mercier, C. Michotey, O. Panaud, N. Picault, N. Rouhier, O. Rué, C. Rustenholz, F. Salin, M. Soler, M. Tarkka, A. Velt, A. E. Zanne, F. Martin, P. Wincker, H. Quesneville, A. Kremer, and J. Salse. **2018**. 'Oak genome reveals facets of long lifespan', Nature Plants, 4: 440-52.

[**P27**] Ranwez, V., E. J. P. Douzery, C. Cambon, N. Chantret, and F. Delsuc. **2018**. 'MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons', Mol Biol Evol, 35: 2582-84.

[**P28**] Dievart, A., C. Gottin, C. Perin, V. Ranwez, and N. Chantret. **2020**. 'Origin and Diversity of Plant Receptor-Like Kinases', Annu Rev Plant Biol, 71: 131-56.

[**P30**] Ranwez, V., Chantret, N., and F. Delsuc. **2021**. 'Aligning Protein-Coding Nucleotide Sequences with MACSE.' In K. Katoh (ed.), Multiple Sequence Alignment. Methods and Protocols. Methods in Molecular Biology 2231, Springer Protocols. Humana Press.

[**P31**] Gottin, C., Dievart, A., Summo, M., Droc, G., Périn, C., Ranwez, V., Chantret, N. **2021**. 'A New comprehensive annotation of leucine-rich repeat-containing receptors in rice'. The Plant Journal, 108(2): 492-508.

Chapitre d'ouvrage

[**P29**] Ranwez, V., and N. Chantret. **2020**. 'Strengths and Limits of Multiple Sequence Alignment and Filtering Methods.' in C. Scornavacca, Delsuc, F., and Galtier, N., editors (ed.), Phylogenetics in the Genomic Era (No commercial publisher. Authors open access book. : The book is freely available at https://hal.inria.fr/PGE.).

Communications

Posters :

[P] Mingeot, D., Chantret, N., de Froidmont, D., Doussinault, G. and Jacquemin, J.M. (**1998**) Search for quantitative trait loci associated with powdery mildew adult resistance in wheat. In Proceedings 50th international symposium on crop protection. Gand. (Mai 1998).

[P] Cenci, A., Chantret, N., Anderson, O., Dubcovsky, J. (**2000**). A half million clone BAC library of durum wheat. Progress report. In International Triticeae Mapping Initiative (ITMI) Public Workshop. University of Delaware. (Juin 2000).

[P] Cenci, A.*, Chantret, N.*, Anderson, O., Dubcovsky, J. (**2001**). A half million clone BAC library of durum wheat. In Plant and Animal Genome IX. San Diego. (Janvier 2001).* The two first authors contributed equally to the work

[P] Chantret, N., Cenci, A., Anderson, O.D., Dubcovsky, J. (**2003**) The analysis of *Triticum monococcum* 101-kb at the *Ha* locus revealed partial conservation of microcolinearity between wheat and rice. In Plant and Animal Genome XI conference. San Diego, Californie, USA. (Janvier 2003).

[P] Haudry, A., Glémin, S., Chantret, N., Cenci, A., Poncet, C., Ravel, C., Brunel, D., Bonnin, I., Hochu, I., Poirier, S. Santoni, S., Bataillon, T., David, J. (**2005**) Evolutionary scenario for the last 12,000 years in wheats. In the 10th International Congress of the European Society of Evolutionnary Biology (ESEB). Jagiellonian University, Cracovie, Pologne (Août 2005).

[P] Haudry, A., Glémin, S., Chantret, N., Cenci, A., Chalhoub, B., Poncet, C., Ravel, C., Brunel, D., Balfourier, F., I., Hochu, I., Poirier, S. Santoni, S., Bataillon, T., David, J. (**2005**) Impact of domestication on the sequence polymorphism of the GSP region in the genome of wheat. In the 4th Plant Genomics European Meetings Congress (Plant GEM). Centre RAI, Amsterdam, Netherlands. (Septembre 2005).

[P] Cenci, A., Chantret, N., Santoni, S., Poirier, S., Gautier, M.F.,Joudrier, P., Bataillon, T., David, J. (**2007**) Comparative analysis at the Ha locus of the two tetraploid sister species, T. turgidum and T. timopheevii. In The Aaronsohn International Triticeae Mapping initiative, April 16-20 2007, Tiberias, Israël.

[P] Ho-Huu J., Ronfort J., De Mita S., Bataillon T., Prosperi J.M., Chantret N. (**2011**) Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.). *Model Legume Congress*, 2011/05/15-19, Sainte-Maxime, France.

[P] Karaki, L., Da Silva, P., Eyraud, V., Gressent, F., Chantret, N., Rizk, F., Rahbé, Y., Royer, C. (**2012**). *Identification of homologous genes to PA1b, a cysteine-rich plant peptide, in Medicago truncatula*. Presented at 3. International Symposium on Antimicrobial Peptides: Today knowledge and future applications, 13-15 juin 2012, Lille, France. http://prodinra.inra.fr/record/185573

[P] Fischer, I., Dufayard, J.-F., Ranwez, V., Chantret, N. (**2012**). Looking for positive selection in recently duplicated genes in plant genomes . Presented at Population Genetics Group “PopGroup” , Glasgow, GBR . http://prodinra.inra.fr/record/183729

[P] Karaki, L., Da Silva, P., Eyraud, V., Gressent, F., Chantret, N., Rizk, F., Rahbé, Y., Royer, C. (**2012**). Identification of homologous genes to PA1b, a cysteine-rich plant peptide, in Medicago truncatula.

Presented at 3. International Symposium on Antimicrobial Peptides: Today knowledge and future applications, Lille, FRA (2012-06-13 - 2012-06-15). http://prodinra.inra.fr/record/185573

[P] Ho-Huu J., Ronfort J., De Mita S., Bataillon T., Prosperi J.M., Chantret N. (**2012**) Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.). SMBE 23-26 juin 2012, Dublin, Ireland.

[P] Fischer, I., Dainat, J., Dufayard, J.-F., Ranwez, S., Chantret, N. (**2013**) Selection positive chez les gènes récemment dupliqués dans les génomes de plantes ; Looking for positive selection in recently duplicated genes in plant genomes. JOBIM 1-4 juillet 2013, Toulouse, France.

[P] Jean-Marie Prosperi, Concetta Burgarella, Nathalie Chantret, Laurène Gay (**2017**). *Using germplasm collections to investigate the genetic architecture of adaptation: a case study in Medicago truncatula.* Presented at : EUCARPIA Genetic Resources 2017 - Crop diversification in a changing world : Mobilizing the green gold of plant genetic resources, Montpellier, France.

[P] Céline Gottin, Anne Dievart, Nathalie Chantret, Vincent Ranwez (**2019**). *Comment annoter et analyser les protéines à motifs répétés : cas des gènes LRR chez A.thaliana et O. sativa ssp japonica.* Presented at : JOBIM 2019 : Journées Ouvertes Biologie, Informatique et Mathématiques, Nantes, France.

[P] Céline Gottin, Anne Dievart, Gaëtan Droc, Nathalie Chantret, Vincent Ranwez (**2020**). *Manual curation and annotation transfer between genomes of LRR-containing genes..* Presented at : JOBIM 2020, Montpellier, France.

Communications orales :

[CO] Chantret, N**.**, Anderson, O., Dubcovsky, J. (**2002**) Comparaison structurale entre *T.monococcum* et *T.durum* dans la région génomique contenant les gènes codant pour les puroindolines, à l'aide de clones BAC. Groupe Recherche Céréales, Nyon, Suisse.

[CO] Chantret, N., Salse, J., Sabot, F., Rahman, S, Bellec, A., Laubin, B., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P., Chalhoub, B. (**2005**) Precise Polyploidy-Related Evolution Mechanisms In *Triticum* Species. In Plant and Animal Genome XIII conference. San Diego, Californie, USA.

[CO] Bataillon, T., DeMita, S., Chantret, N., McKhann, H., Loridon, K ., Santoni, S., Poncet, C., Brunel, D., Delalande, M., Prosperi, J.M., Ronfort, J. (**2009**) Selection footprints in nodulation genes and the prospects for association mapping : insights from a survey of nucleotide polymorphism in *Medicago truncatula*. Model Legume Congress. 12-16 Juin 2009- Asilomar-CA; USA

[CO] Ronfort J, Chantret N, Gay L, De Mita S, Loridon K, Prospéri JM, Siol M and T Bataillon (**2011**). Naturally occurring variation in *Medicago truncatula* : What have we learn from population genetics studies. *Model Legume Congress*, 2011/05/15-19, Sainte-Maxime, France.

[CO] Burgaralla, C., Chantret, N., Gay, L., Prosperi, J.-M., De Mita, S., Young, N., Ronfort, J. (**2012**). Effet des variations climatiques sur la variabilité des gènes déterminant la date de floraison : une étude chez *Medicago truncatula*. In: *Actes du colloque* (p. 54). Presented at 34. Réunion annuelle du Groupe d'Etude de Biologie et Génétique des Populations (Petit Pois Déridé 2012), Avignon, Fance.

[CO] Fischer, I., Dainat, J., Ranwez, V., Glemin, S., Dufayard, J.-F., Chantret, N. (**2014**) Impact of recurrent gene duplication on adaptation of plant genomes; Ecological Genetics Group Meeting ; 14-16 Avril 2014, Newcastle, United Kingdom.

[CO] Chantret N, Ronfort J, Burgarella C, David J, Ecarnot M, Freville H, Gay L, Gouesnard B, Loridon K, Muller MH, Prosperi JM, Ranwez V, Roumet P, Santoni S, Tavaud-Pirra M et Y Vigouroux. (**2014**). Exploiter la diversité génétique pour comprendre les mécanismes de l'adaptation. Journees Scientifiques du Departement BAP, 14-16 avril 2014, Pont Royal, France.

[CO] Pilet-Nayel M-L, Bonhomme M, André O, Hajri A, Boutet G, Badis Y, Chantret N, Ronfort J, Young ND, Baranger A et al. (**2014**). Translational genomics for resistance to Aphanomyces euteiches between Medicago truncatula and pea. In joint conference of the 6 International Food Legumes Research Conference (IFLRC VI) and the 7 International Conference on Legume Genetics and Genomics (ICLGG VII), Saskatoon, Canada.

[CO] Buitink J, Ly Vu J, Neveu M, Dang TT, Ly Vu B, Le Signor C, Ly Vu L, Chantret N, Ronfort J, Prosperi JM et al. (**2021**). Regulation of the plasticity of longevity upon drought in *Medicago truncatula* involves the cryptochrome-interacting bHLH49 transcription factor. In 13th triennial meeting of the ISSS- Seed Innovation Systems for the 21st, Kew, United Kingdom.

ENCADREMENT D'ETUDIANTS

Post-doctorat

Iris Fischer

avril 2012 - juin 2014 « Comparative genomics of gene and gene families » ; financement Agropolis fondation (ARCAD project); co-encadrement JF Dufayard.

juill. 2014 - juin 2016 : financement 'Deutsche Forschungsgemeinschaft' « Detecting footprints of selection at duplicated genes in cultivated and wild flowering plant species using population genetics tools".

Jacques Dainat

jan. 2013 - dec. 2013 : financement Institut Agro Montpellier ; co-encadrement V. Ranwez.

Doctorat

Céline Gottin

oct. 2018 - nov. 2021 : financement CIRAD / Institut Agro Montpellier : « Outils et concepts pour l'annotation de familles de gènes complexes : le cas des récepteurs à LRR chez le riz »

M2

Thibaut Vicat

fev. 2021 - juil. 2021 « Amélioration, mise en forme et distribution d'un pipeline de transfert d'annotation » M2 bioinformatique UM II ; co-encadrement V. Ranwez et C. Gottin.

Enora Gesclin

nov. 2019 - juil.2020 « Elucider les relations phylogénétiques entre *Dioscorea alata* et les potentielles ignames apparentées d'Asie et d'Océanie » M2 bioinformatique Univ. Rennes ; co-encadrement H. Chaïr et V. Ranwez.

Asya Martirosyan

mars 2015 - sept. 2015 « Defensin genes evolution and zinc tolerance in *A. halleri* » M2 UMII parcours MEME ; co-encadrement V. Ranwez.

Maxime De Sario

jan. 2014 - juin 2014 « Polymorphisme de séquence et histoire démographique de la domestication de la luzerne (*Medicago sativa* L.) : approche par Approximate Bayesian Computation. » M2 UMII parcours BEE ; co-encadrement M.-H. Muller.

Emilie Roux

jan. 2009 - juin 2009 « Recherche de traces de sélection sur des gènes impliqués dans la phénologie de la floraison chez *Medicago truncatula* » Master UMII parcours DEPS ; co-encadrement J. Ronfort.

Joan Ho-Huu

jan. 2008 - juin 2008 « Etude de l'évolution moléculaire de gènes dupliqués au sein du genre *Medicago* » Master UMII parcours DEPS ; co-encadrement J. Ronfort.

Saïfallah Bousselmi

fev. 2007 - juin 2007 « Etude de la diversité nucléotidique de *Triticum turgidum.* ssp. » Master UMII Supagro parcours RPIB; co-encadrement A. Cenci.

M1

Fadwa El Khaddar

avril 2022-… « Identification de gène de type NBS-LRR dans le génome comme cibles pour l’édition à la base près » ; Master I Bioinformatique UM II ; co-encadrement P. This et G. Sarah.

Quentin Oliveau

mai 2012 « Modélisation de l’impact du mode de reproduction sur la fixation et le devenir des gènes dupliqués » AgroParisTech 2ème année ; co-encadrement J. Ronfort.

Césure

Fabien Bustos

mars 2022 - … : Institut Agro Montpellier (2ème année) ; co-encadrement G. Sarah.

Audrey Serra

juin 2015 - juil 2015 « Quantification de transcrits alternatifs à l’aide de reads illumina : étude de faisabilité et premiers tests » INSA de Lyon ; co-encadrement V. Ranwez.

TIRES A PART DES PRINCIPAUX TRAVAUX SCIENTIFIQUES

- Gottin, C., Dievart, A., Summo, M., Droc, G., Périn, C., Ranwez, V., and N. Chantret. **2021**. ‘A New comprehensive annotation of leucine-rich repeat-containing receptors in rice’. The Plant Journal, 108(2): 492-508. [**P31**]

- Ranwez, V., A. Serra, D. Pot, and N. Chantret. **2017**. 'Domestication reduces alternative splicing expression variations in sorghum', PLoS ONE, 12: e0183454. [**P24**]

- Fischer, I., A. Dievart, G. Droc, J. F. Dufayard, and N. Chantret. **2016**. 'Evolutionary Dynamics of the Leucine-Rich Repeat Receptor-Like Kinase (LRR-RLK) Subfamily in Angiosperms', Plant Physiol, 170: 1595-610. [**P20**]

- Fischer, I., J. Dainat, V. Ranwez, S. Glemin, J. F. Dufayard, and N. Chantret. **2014**. 'Impact of recurrent gene duplication on adaptation of plant genomes', BMC Plant Biol, 14: 151. [**P18**]

- Ho-Huu, J., J. Ronfort, S. De Mita, T. Bataillon, I. Hochu, A. Weber, and N. Chantret. **2012**. 'Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.)', BMC Evol Biol, 12: 195. [**P15**]

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2021) doi: 10.1111/tpj.15456

A new comprehensive annotation of leucine-rich repeat-containing receptors in rice

Céline Gottin[1,2], Anne Dievart[1,2], Marilyne Summo[1,2], Gaëtan Droc[1,2], Christophe Périn[1,2], Vincent Ranwez[1] and Nathalie Chantret[1,*]

[1]*UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France, and*
[2]*CIRAD, UMR AGAP Institut, F-34398 Montpellier, France*

Received 11 February 2021; revised 23 July 2021; accepted 30 July 2021.
*For correspondence (e-mail nathalie.chantret@inrae.fr).

SUMMARY

Oryza sativa (rice) plays an essential food security role for more than half of the world's population. Obtaining crops with high levels of disease resistance is a major challenge for breeders, especially today, given the urgent need for agriculture to be more sustainable. Plant resistance genes are mainly encoded by three large leucine-rich repeat (LRR)-containing receptor (LRR-CR) families: the LRR-receptor-like kinase (LRR-RLK), LRR-receptor-like protein (LRR-RLP) and nucleotide-binding LRR receptor (NLR). Using LRRPROFILER, a pipeline that we developed to annotate and classify these proteins, we compared three publicly available annotations of the rice Nipponbare reference genome. The extended discrepancies that we observed for LRR-CR gene models led us to perform an in-depth manual curation of their annotations while paying special attention to nonsense mutations. We then transferred this manually curated annotation to Kitaake, a cultivar that is closely related to Nipponbare, using an optimized strategy. Here, we discuss the breakthrough achieved by manual curation when comparing genomes and, in addition to 'functional' and 'structural' annotations, we propose that the community adopts this approach, which we call 'comprehensive' annotation. The resulting data are crucial for further studies on the natural variability and evolution of LRR-CR genes in order to promote their use in breeding future resilient varieties.

Keywords: LRR-receptor-like kinase, LRR-receptor-like protein, nucleotide-binding LRR receptor, annotation curation, pseudogenes, *Oryza sativa*, disease resistance gene.

INTRODUCTION

Modern agriculture is at a critical juncture, as the world's population continues to grow but there is a call to shift away from chemical treatments to deal with current environmental issues. Crop pest and pathogen susceptibility is one of the main causes of annual crop yield loss (FAO, 2018; Savary et al., 2019). Despite an awareness of the harmful environmental impact, massive pesticide use remains a common means to prevent plant diseases today. Studying and understanding plant disease resistance and the underlying evolutionary mechanisms are of utmost importance to make effective widespread use of known sources of resistance through specific breeding programs, while also promoting new resistance engineering for crop sustainability (Bailey-Serres et al., 2019; Tamborski and Krasileva, 2020). The elucidation of resistance mechanisms in plants has highlighted a trove of resistance genes to combat the great and evolving genetic diversity of plant pathogens. The leucine-rich repeat (LRR)-containing receptors (LRR-CRs) are at the forefront of these genes. LRR-CRs share the common structural and functional LRR domain. This domain contains between two and 30+ repetitions of an approximately 24-amino-acid motif, characterized by a conserved skeleton composed mostly of leucine residues (Bella et al., 2008; Kajava, 1998; Kajava, 2012; Matsushima and Miyashita, 2012). These LRR-CRs are classified in three main gene families: LRR receptor-like kinase (LRR-RLK, also named LRR-RK but referred to herein as LRR-RLK), LRR receptor-like protein (LRR-RLP) and nucleotide-binding-site LRR (NBS-LRR or NLR) (Han, 2019; Sekhwal et al., 2015) (Figure 1). The LRR-RLKs and LRR-RLPs are transmembrane receptors composed of an extracellular LRR domain and an intracellular domain. The intracellular domain is a kinase domain for LRR-RLK (Shiu and Bleecker, 2001a; Shiu and Bleecker, 2001b) and a short cytoplasmic tail for LRR-RLP (Fritz-Laylin et al., 2005; Jones and Jones, 1997). Some LRR-RLKs and LRR-RLPs play roles in intercellular communication involved in disease resistance (such as pattern-recognition receptors, PRRs), stress responses or developmental processes (Boutrot and Zipfel, 2017; van der Burgh and Joosten, 2019). Other LRR-RLKs and LRR-RLPs also act as co-receptors or regulators in these signaling pathways (Couto and Zipfel, 2016). NLRs are intracellular receptors composed of a central nucleotide-binding domain (NB-ARC domain) followed by the LRR domain (Burdett et al., 2019; Sekhwal et al., 2015; Tamborski and Krasileva, 2020; Xiong et al., 2020). These proteins can contain other functional domains, such as the toll/interleukin receptor (TIR) domain, the coiled-coil (CC) domain or the resistance to powdery mildew 8 (RPW8) domain, located upstream of the NB-ARC domain.

Over the past few decades, advances in sequencing have provided the research community with an ever-increasing number of complete genomes. These resources have made it possible to revisit gene evolution at the level of entire families and on different evolutionary timescales. LRR-CR genes have been inventoried in many angiosperm genomes, and their numbers have also been compared in a phylogenetic framework to shed light on their evolutionary dynamics (for just some of the more recent articles, see Andersen et al., 2020; Furumizu and Sawa, 2021; Hosseini et al., 2020; Lee et al., 2021; Man et al., 2020; Prigozhin and Krasileva, 2021). A large proportion of LRR-CR genes are thought to evolve through a so called birth-and-death model (McDowell and Simon, 2006; Michelmore and Meyers, 1998; Nei and Rooney, 2005; Richter and Ronald, 2000). In this model, the gene copy number expands by recurrent duplication events and duplicated copies can then follow different evolutionary pathways, such as keeping the original function, acquiring a new function (neofunctionalization) or, more frequently, undergoing a nonfunctionalization process by accumulating nonsense mutations (Innan and Kondrashov, 2010; Leister, 2004). This model explains why LRR-CR genes are found in multiple copies, often organized in large gene clusters, with some genes no longer being functional (Meyers et al., 2003; Mizuno et al., 2020).

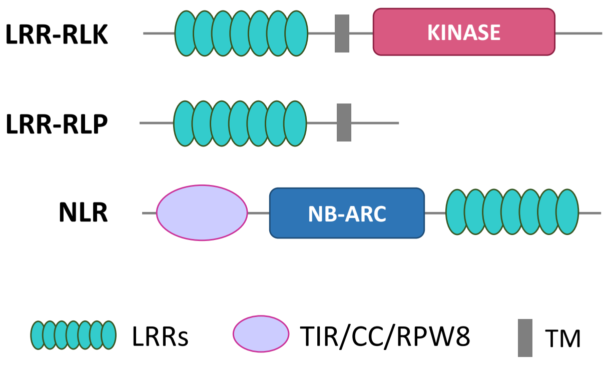


Figure 1. Schematic protein structure of the three LRR-CR subfamilies: LRR-RLK, LRR-RLP and NLR. TM, transmembrane domain; CC, coiled coil domain; TIR, toll-interleukin receptor; RPW8, resistance to powdery mildew 8 domain.

Comparative genomic studies have led to considerable progress in understanding the evolutionary dynamics of LRR-CR gene families, but these studies are highly dependent on the accuracy of annotation procedures. Given the increasing avalanche of sequence data, the most reasonable approach is to rely on automatic annotation. Gene and protein sequence annotation are thus crucial and the target of considerable effort. Structural gene annotation is geared towards identifying coding sequences within genomic data and documenting the associated gene features (e.g. introns, exons and untranslated regions, UTRs) (Wilming and Harrow, 2009). The most widely used structural annotation pipelines, such as the Ensembl pipeline for gene annotation (Aken et al., 2016), Augustus (Stanke and Waack, 2003) and Gnomon (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/), rely on: (i) *ab initio* gene structure determination according to rules learned on pre-existing annotations; and/or (ii) comparative approaches, i.e. using sequence homology with available RNAseq data and/or with a closely related annotated genome. Those methods allow large-scale studies with standardized approaches, yet they are not completely reliable, especially for complex multigene families. Indeed, repetitions are known to impair gene annotations (Bayer et al., 2018; Fawal et al., 2014) and there are also genome assembly issues (Torresen et al., 2019). The difficulty is twofold in the case of LRR-CRs: several similar genes are present in the genome as a result of gene duplication events, whereas each gene contains several similar motifs because of the repetitive structure of the LRR domain. The automatic annotation and classification of LRR-CRs is thus especially challenging. For example, although multiple studies have reported that there are more than 800 LRR-CR loci in the rice variety Nipponbare, the number of genes per family is variable, e.g. 374–498 NLR proteins (Li et al., 2010; Li et al., 2016; Shao et al., 2016; Stein et al., 2018; Zhou et al., 2004), 292–435 LRR-RLKs (Dufayard et al., 2017; Hwang et al., 2011; Man et al., 2020; Sun and Wang, 2011) and 90 LRR-RLPs (Fritz-Laylin et al., 2005). These variations are to a large extent linked to the annotation version chosen for the analysis and to the decision rules for gene detection and classification. Scientists sometimes perform the manual curation of gene annotations to limit these uncertainties and achieve high-quality comprehensive analyses, as in the case of Arabidopsis and *Solanum lycopersicum* (tomato) NLR genes (Jupe et al., 2013; Meyers et al., 2003; Van de Weyer et al., 2019) or *Oryza sativa* (rice) Nipponbare LRR-RLK genes (Sun and Wang, 2011).

Rice was the first monocotyledon plant to have its genome entirely sequenced and three different annotations of its reference genome, *O. sativa* ssp. *japonica* cv. Nipponbare (Kawahara et al., 2013), are currently available: one from the Michigan State University Rice Genome Annotation Project (MSU, http://rice.uga.edu) (Yuan et al., 2003), one from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) and the current reference genome from the Rice Annotation Project of the International Rice Genome Sequencing Project (IRGSP, https://rapdb.dna.affrc.go.jp) (Sakai et al., 2013). We first implemented the LRRPROFILER pipeline to compare them with regards to the LRR-CR protein repertoire. This program builds subfamily- and genome-specific LRR hidden Markov model (HMM) profiles, detects LRR-CR proteins that contain LRR motifs and accurately locates LRR motifs within these proteins. We ran the LRRPROFILER pipeline in parallel on the three rice predicted proteomes and found that they greatly differed in terms of the number of LRR-CR genes and their structural annotations. We therefore performed a manual curation of the whole Nipponbare LRR-CR repertoire annotation. To do so, for gene models that diverged between the three annotations, we looked for the reasons of divergence and decided, when appropriate, to supplement the gene models with sequence fragments undoubtedly derived from LRR-CR-encoding genes. In turn, we provided objective information, i.e. whether the gene models were canonical or non-canonical. To be qualified as canonical a gene model had to fulfil all of these conditions: presence of a start codon; presence of a terminal stop codon; absence of an in-frame stop codon; absence of frameshifts; and absence of unexpected intron splicing sites. Conversely, any gene violating at least one of these constraints was qualified as non-canonical. Finally, we also propose a strategy to transfer these manually curated LRR-CR gene annotations to Kitaake, the closest related japonica genome that has been sequenced (Jain et al., 2019). We then analyzed the observed variations in gene numbers and LRR motifs between Nipponbare and Kitaake genotypes while using the available automatic annotations and our manually curated annotations (hereafter referred to as 'comprehensive'). This comparison demonstrated how erroneous conclusions can readily be drawn when relying solely on automatic structural and functional annotations for this complex gene family. The curated comprehensive LRR-CR annotation introduced in this article is available online through a dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc).

RESULTS

Inconsistencies among three publicly available Nipponbare rice LRR-CR annotations

We used LRRPROFILER, a newly developed pipeline (see Experimental procedures and Data S1, Methods S1, Figures S1 and S2 and Table S1 for LRRPROFILER validation results, performed on a manually reviewed *Arabidopsis thaliana* protein data set and on the whole Arabidopsis proteome, including a comparison with the LRRPREDICTOR tool; Martin et al., 2020), to identify, annotate and classify into gene subfamilies the LRR-CR protein sequences of the three publicly available Nipponbare proteomes (MSU, IRGSP and NCBI). The total number of LRR-CRs identified varied markedly according to the annotation: we identified 1226 LRR-containing sequences in the MSU predicted proteome, 1047 in that of IRGSP and 1073 in that of NCBI (Table 1). The distribution patterns of these proteins in the different subfamilies also varied according to the annotations. For instance, the number of predicted genes fluctuated less for the LRR-RLP subfamily than for the NLR subfamily, for which 60% more NLRs were detected in the MSU proteome (418 proteins) compared with the IRGSP proteome (282 proteins). For comparison, we conducted a similar analysis on nine transcription factor (TF) subfamilies, for which we assumed that the annotation process would be easier as they had a more conserved structure and, although having undergone expansion events, were not evolving under a birth-and-death model (Lai et al., 2020). The TF data set contained between 874 and 1041 genes, according to the annotations, and this number was similar to that of LRR-CR. To assess whether the identified genes were at the same genomic location or not, we measured the overlap of the three predicted gene sets. The percentage of loci for which a gene model was present in all three annotations was 52.3% for LRR-CR genes and 69.5% for TF (Figure 2a), indicating that the three annotations were more congruent for TF genes. Moreover, the percentage of loci in which only one annotation detected a gene was 19.5% for LRR-CR genes, compared with only 12% for TF genes.

Even when a gene was predicted by the different annotations, the predicted structure of the gene sometimes varied between predictions. One way to address this issue is to compare the length of the predicted proteins for genes positioned at the same locus. Note that this is a conservative approach. Indeed, although a predicted protein length difference between two gene models indicated that the gene models differed, the reverse was not true, as identical

Table 1 Number of LRR-CR sequences in the predicted proteomes from three publicly available annotations for the Nipponbare rice reference genome. Sequences were identified and classified into subfamilies using the LRRPROFILER pipeline

	Total	LRR-RLKs	LRR-RLPs	NLRs	Others[a]
IRGSP	1047	237 (22.6%)	160 (15.3%)	282 (26.9%)	368 (35.1%)
MSU	1226	329 (26.8%)	141 (11.5%)	418 (34.1%)	338 (27.6%)
NCBI	1073	305 (28.4%)	121 (11.3%)	361 (33.6%)	286 (26.7%)

[a]F-box-LRR and unclassified (UC) sequences.

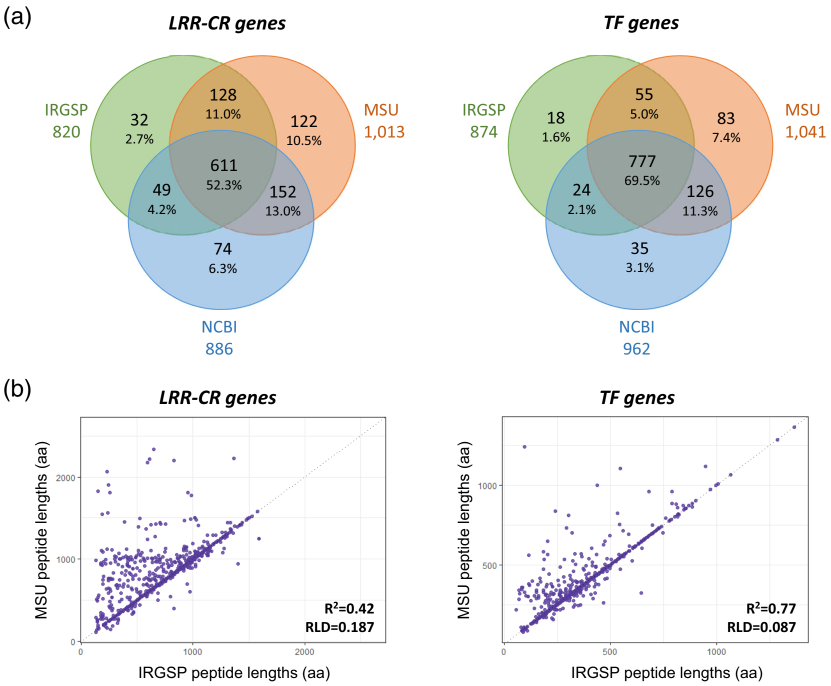


Figure 2. Comparison of publicly available MSU, IRGSP and NCBI annotations for the Nipponbare rice reference genome for two types of genes: LRR-containing receptors (LRR-CRs) and transcription factors (TFs).

(a) Venn diagrams representing the number of overlapping gene models for LRR-CRs and TFs among the MSU, IRGSP and NCBI annotations. To be considered as overlapping, gene models from two (or three) different annotations should have at least one nucleotide in common (overlapping loci). The total number of genes in each annotation does not correspond to the total number in Table 1 because of the complex relationships between loci: for instance, a single NCBI gene can overlap with a gene in IRGSP and another in MSU, whereas these IRGSP and MSU genes do not overlap.

(b) Dot plots representing the polypeptide length in amino acids (aa) for genes predicted by both IRGSP and MSU annotations. On the left, LRR-CRs and on the right, TFs. The R^2 and the average relative length difference (*RLD*) values are given at the bottom right for each gene family. For all pairwise comparisons among IRGSP, MSU and NCBI, see Figure S3.

predicted protein lengths did not guarantee that the gene models were identical. A comparison of predicted protein lengths for all LRR-CR gene pairs located at the same locus but predicted by two different annotations is presented in Figure 2b and Figure S3. Here, again, the number of genes with a difference in the predicted protein length highlighted a substantial annotation discrepancy. This difference was greater for LRR-CR genes than for TF genes. As an example, when IRGSP and MSU were compared, the average difference between the predicted product sizes was 18.7% for LRR-CR loci (with an R^2 of only 0.42), whereas this difference was only 8.6% for TF loci (with a much higher R^2 of 0.77) (Figure 2b). These results highlight the extent to which annotations generally differ, but more particularly for LRR-CR gene subfamilies. These comparisons also showed that LRR-CR genes predicted by IRGSP were generally shorter than those predicted by MSU or NCBI at the same locus (Figure 2b and Figure S3).

Manual re-annotation of LRR-CR-encoding loci in the Nipponbare rice genome

Here we provide a brief description of the procedure that we followed to manually curate LRR-CR annotations (Figure 3a). First, note that for the sake of traceability the procedure retained one of the three proposed gene models as much as possible. For a given locus, we first selected one of the gene models among the available annotations based on the completeness of the predicted protein. We then applied our expertise to the selected gene model by combining protein and nucleotide data. At the protein level, we checked that all of the expected domains for each subfamily were present (e.g. LRRs and TM for LRR-RLPs and LRR-RLKs, kinase for LRR-RLKs or NB-ARC for NLRs) in the right order, with the expected length and interdomain intervals. Protein domain information was particularly useful for detecting potential gene fusion and fission. At the nucleotide level, we examined: (i) whether the gene models had the expected intron/exon structure (e.g. introns, when present, are often found at the same exact position); (ii) whether nearby open reading frames (ORFs) belonging to LRR-CR-encoding sequences were present; and (iii) whether the gene models included suspicious introns, such as short introns, enabling the gene to sidestep stop codons or frameshifts, especially when they were never found in homologs (Figure 3b). Any structural annotation containing an in-frame stop codon or a frameshift (i.e. any gap in coding sequence that was not an intron but that changed the translation phase), lacking a start codon or a terminal stop codon, or presenting an unexpected splicing site (different from the GT-AG and GC-AG donor/acceptor canonical splicing sites) was called ‘non-canonical’. This careful inspection was facilitated by viewing the sequence annotations with the ARTEMIS editor (Carver et al., 2012).

In a last step, we also looked for LRR-containing sequences that would have been missed by the three publicly available annotations. The Nipponbare reference genome was split into 1-kb segments with overlapping 100-bp borders, translated into amino acid sequences in the six reading frames (as performed by Steuernagel et al., 2020), and domains (LRR, kinase, NB-ARC, etc.) were searched

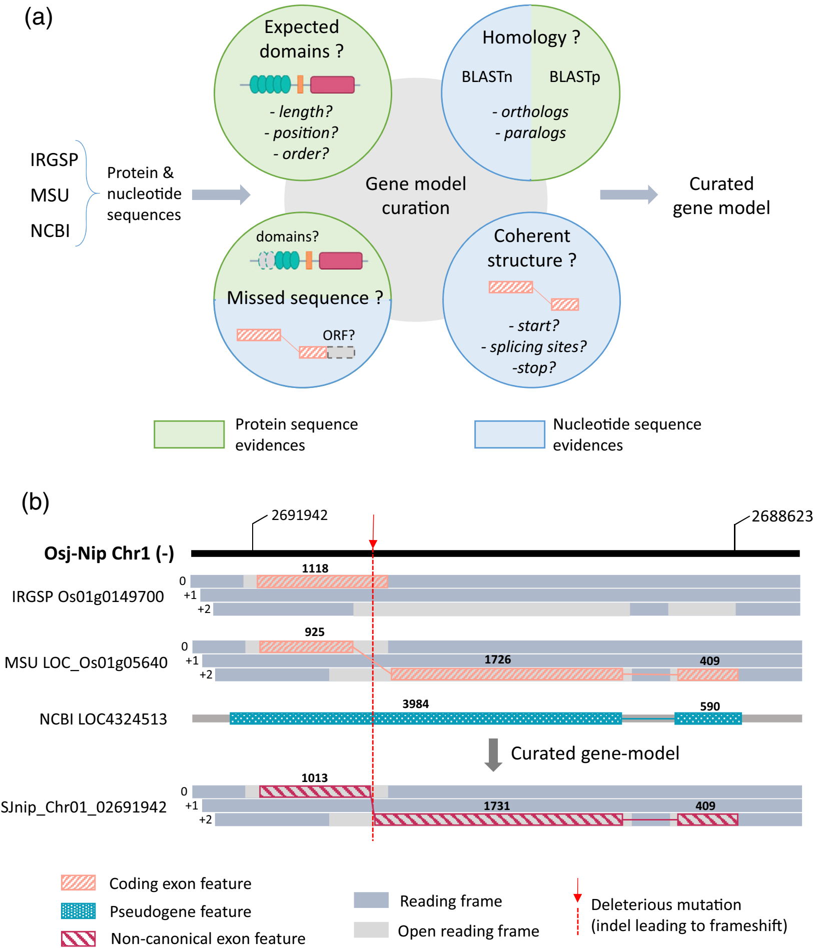


Figure 3. Manual curation of LRR-CR gene model strategy and example of annotation inconsistencies.

(a) Schematic representation of the strategy used to curate Nipponabre LRR-CR gene models. An initial gene model was selected from the three public annotations. This gene model was gradually modified based on protein and nucleotide sequence evidence. The curated model was then classified as canonical or non-canonical.

(b) Schematic representation of an example of inconsistency between gene models from publicly available annotations and how the curation was performed. The gene is an LRR-RLK located on chromosome 1 of the Nipponbare genome. The numbers above the boxes indicate the length of the feature. In this example, an indel mutation caused a frameshift in the first exon of the gene. The IRGSP annotation retrieved the first part of the coding sequence, stopping at the first stop codon on frame 0. The MSU annotation retrieved a longer coding sequence but sidestepped the indel mutation by introducing a 'dubious' intron in order to reach the open reading frame (ORF) on the +2 frame. This 'dubious' intron was abnormally short and contained a sequence highly homologous to the coding sequence in other paralogous gene copies. The NCBI annotation gave a pseudogene feature, i.e. a feature from which a protein sequence could not be deduced: the cDNA sequence is available but would not allow protein translation as it would be in the wrong reading frame after the mutation. The curation took advantage of the three annotations. It retried a cDNA sequence that overlapped the complete former coding sequence in the two successive correct reading frames via the identification of the indel mutation. The identification of the indel mutation was clear cut as the gene was tagged as 'non-canonical' with the presence of a frameshift, but it allowed a complete protein sequence to be deduced and used for sequence comparison and alignment.

with HMMSEARCH. All the results were concatenated and filtered for redundancies. The retained new sequences of interest had to contain at least three LRR motifs in tandem. If another domain was detected at less than 5 kb from these LRR motifs, the sequence of interest was enlarged to also include these domains. These sequences, not overlapping known LRR-CR exons, were compared with other plant genomes using BLAST to screen for the potential presence of a gene model in the region under consideration.

The final set of manually validated LRR-CR loci on the Nipponbare genome consisted of 1058 genes (350 LRR-RLKs, 147 LRR-RLPs, 503 NLRs and 58 UCs) (Data S2; Table 2). Among these 1058 genes, eight (one LRR-RLK, three LRR-RLP and four UC) were located at loci for which none of the three publicly available annotations detected a gene. The LRR-RLK was a canonical full-length sequence on the forward strand of chromosome 2 (from 6 831 702 to 6 834 761). Note that this sequence is actually present in GenBank under accession number EAZ22278.1 and is located on the reverse strand in a non-coding region of the *Os02g0222500* gene. The other seven are non-canonical truncated genes. In addition, for seven of these 1058

Table 2 Number of LRR-CR proteins in the predicted proteomes from our curated annotations for the Nipponbare rice reference genome. Sequences were identified and classified into subfamilies using the LRRPROFILER pipeline

	Total	LRR-RLK	LRR-RLP	NLR	UC
LRR-CR loci[a]	1058 (8)	350 (1)	147 (3)	503 (0)	58 (4)
Modified loci (%[b])	328 (31.0%)	56 (16%)	55 (37.4%)	197 (39.2%)	20 (34.5%)
Non-canonical loci (%)	306 (28.9%)	53 (15.1%)	48 (32.7%)	183 (36.4%)	22 (37.9%)
Modified and non-canonical (%)	274 (25.9%)	43 (12.3%)	43 (29.3%)	170 (33.8%)	18 (31.0%)

[a] Numbers in parentheses are newly identified LRR-CR genes.
[b] Percentages were calculated based on the number of manually curated genes, i.e. the total number of genes minus the number of newly identified genes.

validated genes, the LRRPROFILER pipeline did not detect any further LRR motifs in the predicted protein. LRR motifs were initially detected for these genes, but at the threshold limit when using HMM profiles built on the basis of the initial data set (for details, see the LRRPROFILER pipeline section in the Experimental procedures). When using the slightly different HMM profiles obtained with the final data set, the same LRR motifs were no longer detected as they did not surpass the threshold. However, a careful manual inspection showed that the LRR domain was present but contained divergent LRR motifs, thereby complicating the automatic detection. Consequently, these genes were kept and classified according to the presence of the other domains (kinase or NB-ARC). These seven genes included one LRR-RLK and six NLRs.

Among these 1058 LRR-CR genes, 328 (197 NLR, 56 LRR-RLK, 55 LRR-RLP and 20 UC) were manually modified because none of the three publicly available annotations had a satisfactory gene model based on the previously defined criteria (Data S2; Figure 3a). The overall proportion of modified loci was 31.0% (328/1058), and varied markedly according to the gene subfamily considered. Only 16% of LRR-RLK loci were modified, whereas 37.4% of the LRR-RLP loci and 39.2% of the NLR loci were modified (Table 2). Among these 1058 LRR-CR genes, 306 (28.9%) were non-canonical. Again, the different gene subfamilies did not contain the same proportion of non-canonical gene models. Very similar to what was observed regarding the proportion of modified gene models according to gene subfamily, non-canonical gene models concerned only 15.1% of the LRR-RLKs, compared with 32.7 and 36.4% of the LRR-RLPs and NLRs, respectively. Thus, 274 genes were both non-canonical and modified, representing 83.5% of the total modified loci (274 over 328) and 89.5% of the non-canonical loci (274 over 306) (Table S2). The remaining 32 non-canonical genes were either unreported by any of the annotations (seven) or were reported by the NCBI as pseudogene or gene models having putative errors in the genomic sequence (25, see below).

One way to assess the relevance of our expert LRR-CR annotation is to compare the number of functional domains (TMs, NB-ARCs, kinases and LRRs) found in LRR-CR proteins derived from the reference annotations to the number of functional domains found in the proteins derived from our expert annotation (Figure S4). These comparisons revealed that quite a few more LRR-CR domains were found in our manual annotation as compared with the publicly available annotations. For example, when compared with the reference IRGSP annotation, our expert annotation highlighted 29% more TM, 42% more NB-ARC, 33% more kinase and 20% more LRR motifs.

Annotation of the LRR-CR genes in the rice cultivar Kitaake

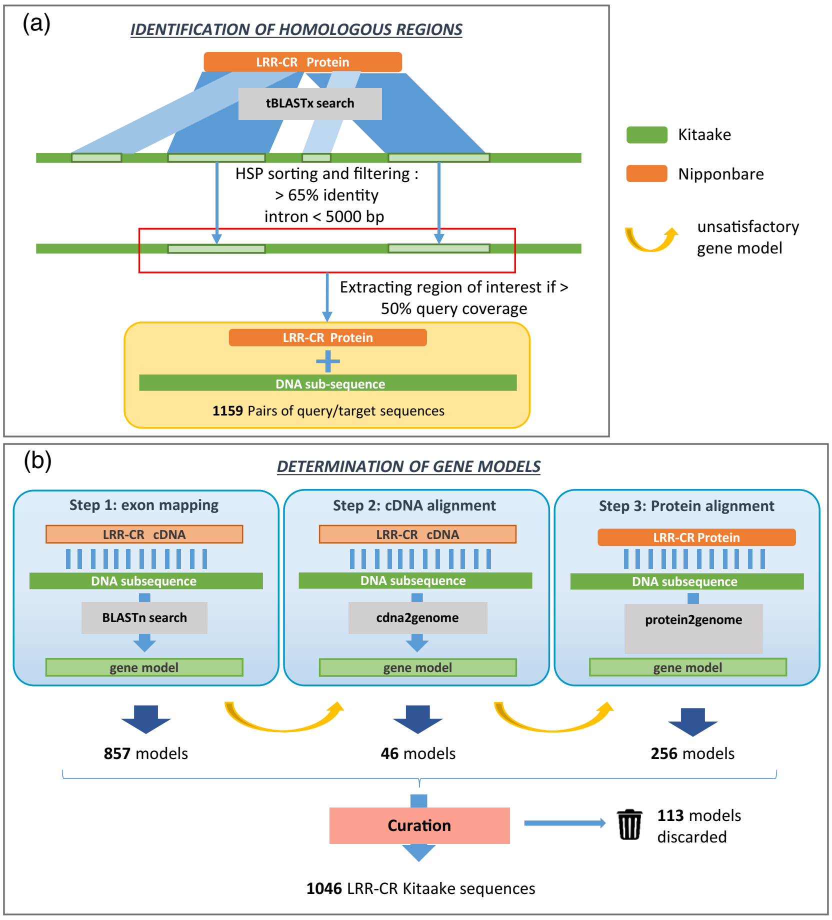
Kitaake is another *O. sativa* ssp. *japonica* variety for which a complete genomic sequence is available (Jain et al., 2019). In order to compare the LRR-CR repertoire between Nipponbare and Kitaake and limit the need for manual curation in the re-annotation of this closely related rice cultivar, we developed a strategy to transfer our expert annotations from the Nipponbare to the Kitaake genome.

The strategy summarized in Figure 4 starts by identifying Kitaake genome regions that are homologous to Nipponbare LRR-CR sequences. Then it successively takes into account three levels of annotation transfer, depending mostly on the level of sequence identity of each considered region with the LRR-CR gene that identified it. At each locus, our strategy strives to retrieve the most probable gene model with the idea that, if possible, it should be canonical. At the end of the process, LRR-CR gene models that are found to be non-canonical or having a dubious protein structure in Kitaake are manually checked and corrected if needed. At this step, the transfer allowed us to identify 1046 LRR-CR genes in the Kitaake genome.

As carried out for Nipponbare, the Kitaake genome was finally scanned with LRR HMM profiles using HMMSEARCH for new LRR-CR identifications. This procedure allowed us to annotate 18 additional genes, thereby leading to a total of 1064 LRR-CR genes in the Kitaake genome.

The LRRPROFILER pipeline was used on the 1064 predicted Kitaake proteins and allowed us to detect LRR in 1053 of them; 999 were further classified into a LRR-CR subfamily and 54 remained in the UC group. The automatic detection of LRR failed for 11 genes. As carried out for Nipponbare,

Figure 4. Schematic representation of the annotation transfer strategy between closely related genomes.
(a) Identification of LRR-CR homologous regions. Nipponbare LRR-CR proteins were used to search regions of interest in the Kitaake genome using tBLASTx. BLAST hits with over 65% identity were ranked in the LRR-CR query protein sequence order and used to define the region boundaries. If the filtered BLAST hits within a Kitaake region covered more than 50% of the query LRR-CR sequence, then the Kitaake sequence of this region was extracted and linked to the Nipponbare LRR-CR query protein.
(b) Determination of gene models. The process strives to give a gene model for each region of interest identified in the Kitaake genome. The annotation is attempted in three consecutive steps. If the model from one step is unsatisfactory, i.e. gives an alignment of poor quality with the Nipponbare query protein, the process goes to the next step for this region. At the end of the third step, gene models that remained unsatisfactory were manually checked. This process allowed us to annotate 1046 genes in the Kitaake genome.



at this step, manual validation of the protein annotations confirmed the presence of an LRR domain of the expected size and located at the expected positions. These 11 genes were therefore kept in the final data set. The gene subfamilies for these 11 loci were determined based on other functional domains and a homology search against other LRR-CR protein sequences. Finally, the LRR-CR gene set from Kitaake was composed of 360 LRR-RLKs, 140 LRR-RLPs, 510 NLRs and 54 UCs (Data S3; Figure 5). These numbers were very similar to those obtained for Nipponbare, i.e. 350 LRR-RLKs, 147 LRR-RLPs, 503 NLRs and 58 UCs.

We then tagged all of these Kitaake LRR-CR gene models as either canonical or non-canonical. We obtained 742 (69.7%) canonical genes and 322 (30.3%) non-canonical genes. Again, the proportions of canonical and non-canonical genes per subfamily for Kitaake were very similar to those obtained for Nipponbare (Figure 5).

A notable result is that our strategy enabled us to identify 114 LRR-CR genes (48 of which were canonical) that were not present in the publicly available annotation of the Kitaake genome: 17 LRR-RLKs, 24 LRR-RLPs, 50 NLRs and 23 UCs.

All LRR-CR loci annotation and sequence data for the Nipponbare and Kitaake genomes can be viewed and downloaded on the dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc).

Comparison of LRR-CR allelic pairs between Nipponbare and Kitaake

Nipponbare and Kitaake are two varieties of the same subspecies: *O. sativa* ssp. *japonica*. As such, for the majority of the genes found in Nipponbare, an allele (i.e. a version of the same gene located at the same chromosomal location) was expected to be found in Kitaake. By using SYNMAP (Lyons and Freeling, 2008), we identified 1002 allelic pairs (representing 90.5% of the total number of loci) between Nipponbare and Kitaake (Data S4). In addition, we noticed that for three NLR gene pairs located close to each other on chromosome 9 in the Nipponbare genome, three consecutive genes on chromosome 3 of the Kitaake genome were found with 100% identity with regards to their predicted coding sequence. The intergenic sequences of these two regions also had a high level of identity (99% over 40.5 kb), suggesting that these three genes are located in a translocated region of the genome.

First, to assess the impact of re-annotations on the number of LRR motifs in the alleles, the number of LRR motifs

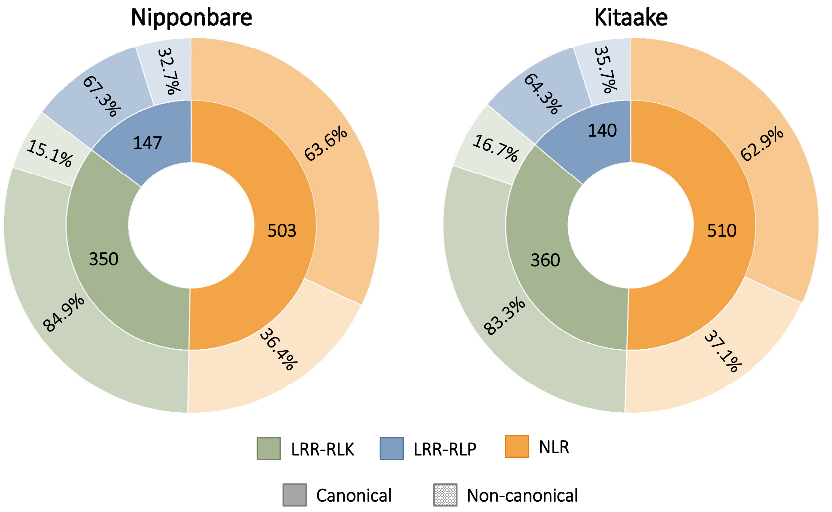


Figure 5. Proportion of canonical and non-canonical loci per gene subfamily in our Nipponbare and Kitaake expert annotations. Percentages were calculated per gene subfamily. The inner circle provides the number of loci per family, with a different color for each. The outer circle shows a lighter/darker version of the loci family color to represent the fraction of the non-canonical/canonical members, respectively, within this gene family.

predicted in Nipponbare was compared with the number of LRR motifs predicted in Kitaake for each pair of allelic sequences. To obtain a precise annotation of the LRR motifs in each protein, we used the LRRPROFILER pipeline. The same procedure was also applied on allelic pairs identified between the publicly available annotations of Nipponbare and Kitaake. We observed a mean difference in LRR number per protein of 3.58 when comparing the publicly available annotations (IRGSP for Nipponbare and the only one that exists for Kitaake) (Figure 6 and Figure S5). This difference fell to 0.6 when our re-annotated data were compared. Using our curated annotations hence led to LRR number predictions that were much more consistent between Nipponbare and Kitaake alleles, and this trend was observed for all LRR-CR gene subfamilies. Moreover, the mean difference in LRR number still varied between LRR-CR gene subfamilies, with greater conservation of LRR motif numbers between LRR-RLK and LRR-RLP alleles than between NLR alleles.

Second, we analyzed the re-annotated allelic pairs related to their canonical or non-canonical status. Among the 1005 pairs (1002 allelic plus three translocated pairs), 688 (68.5%) were pairs of canonical gene models, 269 (26.8%) were pairs of non-canonical gene models and 48 (4.8%) were pairs of genes found to be canonical in only one of the two cultivars. Interestingly, 83.1% of the LRR-RLK pairs were canonical in both cultivars, compared with only 63.9% of the LRR-RLP pairs and 60.3% of the NLR pairs (Table 3).

To go further into this comparison, for each of the 1005 pairs of LRR-CR alleles, the fraction of exact matches along the cDNA pairwise global alignment (i.e. their percentage of identity) was computed. This cDNA identity was about 98.6% on average. The highest identity rate (99.3%) was obtained for alleles belonging to the LRR-RLK subfamily, followed by the NLR (98.2%) and LRR-RLP (97.9%) subfamilies. On average, non-canonical conserved gene pairs (NC/NC category in Table 3) had a lower identity level (97.9%) than conserved canonical gene pairs (99.4%). The lowest level of sequence identity (91.2%) was noted between gene pairs with one cultivar having a canonical form and the other cultivar having a non-canonical form (categories C/NC and NC/C in Table 3). Only 25 pairs of alleles (three LRR-RLK, four LRR-RLP, 17 NLR and one UC) shared less than 80% cDNA identity as a result of both deletions (up to 1.7 kb) and high sequence divergence. Two of these NLRs are located in the RGA5 and Pik clusters that both hold resistance genes to rice blast disease (Table S3) (Li et al., 2007; Okuyama et al., 2011).

Genotype-specific LRR-CR genes in Nipponbare and Kitaake genomes

This gene presence–absence variation (PAV) analysis revealed that 48 LRR-CR genes were present only in Nipponbare and 58 LRR-CR genes were present only in Kitaake, of which 30 (six LRR-RLK, nine LRR-RLP, 13 NLR and two UC) and 34 (11 LRR-RLK, three LRR-RLP and 20 NLR), respectively, were canonical. Note that among the 11 LRR-RLK Kitaake-specific genes are the two *Xa21* transgenes introduced into the KitaakeX sequenced genome (Jain et al., 2019). The *Xa21* gene was initially cloned from the wild rice species *Oryza longistaminata* (Song et al., 1995). We indeed identified these two transgenes at positions 28 161 378 and 28 165 947 on chromosome 6, in accordance with published data (Jain et al., 2019). Among the Nipponbare-specific genes, two LRR-RLK (OsLP2 and RLCK354), three NLR (RPR1, STA260 and Osh359-3) and one UC (Bph33) have been named previously (Hu et al., 2018) (Sakamoto et al., 1999; Thilmony et al., 2009; Yao et al., 2018).

The genotype-specific genes were not evenly distributed on the genomes. Most of them, 72.9% (35/48) and 60.3%

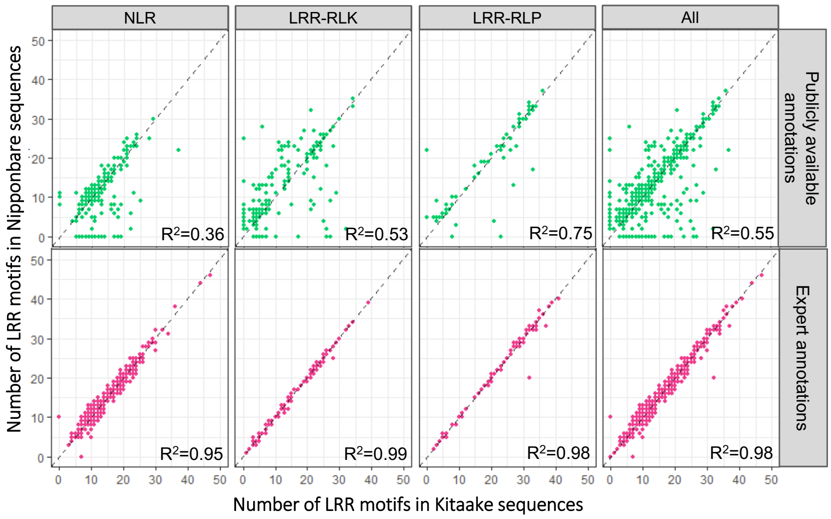


Figure 6. Comparison of LRR motif numbers between Nipponbare and Kitaake LRR-CR alleles, according to the annotations used. In green, comparison between two publicly available annotations of Nipponbare and Kitaake using IRGSP reference data for Nipponbare. In pink, comparison between our Nipponbare and Kitaake expert annotations.

Table 3 Number of allelic pairs between Nipponbare and Kitaake cultivars according to categories and subfamilies

Allele categories	Total	LRR-RLK	LRR-RLP	NLR	UC
C/C[a]	688 (68.5%)	285 (83.1%)	85 (63.9%)	289 (60.3%)	29 (58.0%)
NC/NC[a]	269 (26.8%)	47 (13.7%)	41 (30.8%)	163 (34.0%)	18 (36.0%)
C/NC[a]	29 (2.9%)	7 (2.0%)	4 (3.0%)	15 (3.1%)	3 (6.0%)
NC/C[a]	19 (1.9%)	4 (1.2%)	3 (2.3%)	12 (2.5%)	0 (0%)
Total	1005	343	133	479	50

[a] The four categories partitioned the loci according to whether they were canonical (C) or non-canonical (NC) in Nipponbare/Kitaake. Numbers in parentheses are the percentages per subfamily.

(35/58) of the Nipponbare- and Kitaake-specific loci, respectively, were located on chromosomes 2, 11 and 12. On these chromosomes, some gene clusters were entirely composed of genotype-specific genes (Figure 7a). Other genotype-specific genes were found dispersed in regions containing conserved allelic pairs (Figure 7b). Chromosome 11, which also contained about a fifth of all LRR-CR genes, hosted 43 of the 106 (40.6%) cultivar-specific loci.

Moreover, for more than half of these canonical genes (38 out of 64) the highest homology found in the Nipponbare or the Kitaake proteome is <80% of identity. Note that among these 38 genes, five Kitaake genes and seven Nipponbare genes have more than 95% of identity with indica cultivar proteins. Thus, the divergence of these genotype-specific proteins, related or not to the breeding histories of these varieties, highlights the variability of the LRR-CR repertoires between these two closely related accessions.

Finally, we took advantage of having the LRR-CR repertoire for both Nipponbare and a second rice genome to quantify putative sequence errors in the Nipponbare assembly. Among the 306 Nipponbare non-canonical genes, 241 (78.8%) presented at least one nonsense mutation also found within the Kitaake allele. These mutations are then assumed to be real. The remaining 65 non-canonical genes were manually checked and four different cases were identified: (i) Nipponbare-specific genes (18 genes, 27.7%); (ii) Kitaake allelic canonical genes (19 genes, 29.2%); (iii) Nipponbare genes that have been classified as non-canonical for a different reason than the non-canonical Kitaake allele (six genes, 9.2%); and (iv) genes for which the 'RefSeq' data of NCBI reported a potential sequence error in Nipponbare (22 genes, 33.8%). The sequence of these 65 genes was compared using BLASTn with a full-length complementary DNA (FLcDNA) clone library (Rice Full-Length cDNA Consortium, 2003) and with 14 Illumina sequence read archives (SRAs) of Nipponbare (Wang et al., 2018). The mutations observed in 18 and 40 genes (89.2%) were validated by the FLcDNA library and SRA, respectively. For four genes, no hits were obtained (6.2%). For the three remaining genes, a genomic sequence error was detected. The first one contained an 'N' that generated a frameshift, the second one had a small inversion of 24 bases that turned out to be erroneous and the third one had a wrong indel. These three genes, belonging to the NLR subfamily, were tagged in the data sets. These genes have not yet been described in the literature.

DISCUSSION

In recent years, in the wake of the gigantic volume of genome sequenced data available, it was exciting to undertake

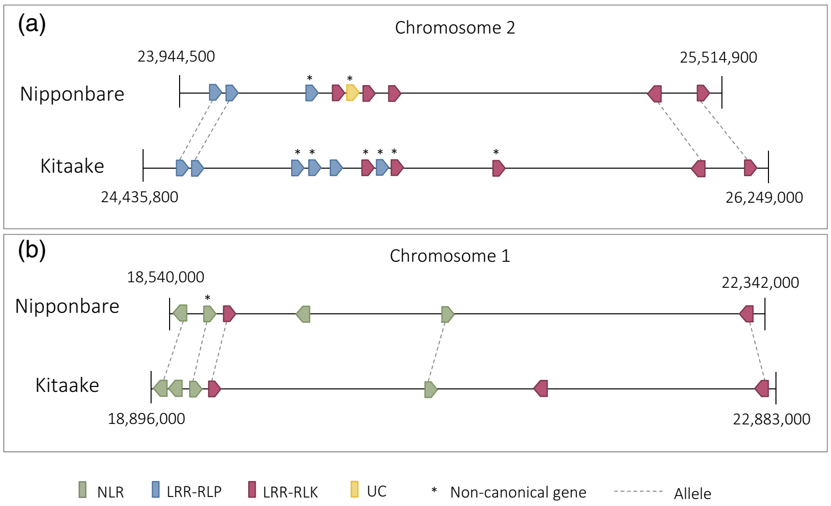


Figure 7. Schematic representation of two large loci on chromosomes 1 and 2 containing cultivar-specific LRR-CR genes.
(a) Representation of an unconserved cluster between Nipponbare and Kitaake on chromosome 2. Five and seven genes in Nipponbare and Kittake, respectively, were cultivar specific. The unconserved region was framed by four conserved genes, i.e. two LRR-RLPs and two LRR-RLKs.
(b) Representation of a conserved region between Nipponbare and Kitaake on chromosome 1 hosting cultivar-specific genes. The Nipponbare region hosted a cultivar-specific NLR, whereas the corresponding Kitaake region hosted two cultivar-specific genes, i.e. an NLR and an LRR-RLK.

evolutionary studies of gene families. We have been part of this collective enthusiasm, and like many others have based our research conclusions on perfectible versions of automatic structural and functional gene annotations (Dufayard et al., 2017; Fischer et al., 2016). Although previous genome-wide phylogenetic approaches on LRR-CR gene families enhanced our knowledge on their evolution, they almost never included a data curation step. Indeed, the manual re-annotation of gene families is a laborious time-consuming task, especially when dealing with large complex gene families, such as LRR-CR, and even more so when dealing with many plant genomes. Despite that automatic annotation tools are continuously improving, and remain essential at the genome level, human expertise is clearly still needed to achieve the level of annotation accuracy suitable for finer and deeper analyses. Today, we have finally undertaken this re-annotation work because we are convinced that these curated data are required to produce new reliable results on the evolution of these gene families, especially in the current pangenomic era. Here, we describe a new so-called ‘comprehensive’ annotation strategy. We hope that this annotation process will gain its place alongside the structural and functional annotations in use so far.

Automatic annotations give inconsistent gene models on complex multigenic families

Our comparison of the three publicly available annotations for the Nipponbare rice reference genome showed major discrepancies regarding the total number of LRR-CR genes, the number of LRR-CRs assigned to each subfamily and the gene models (Figure 2 and Figure S3; Table 1). These differences were greater for LRR-CR genes compared with other genes such as TFs. Automatic annotation pipelines appeared to be suitable overall for many gene families, but they led to a large proportion of inconsistent gene models when annotating complex multigenic families like LRR-CR. The annotation of fast-evolving multigene families is especially challenging for automatic approaches because a high duplication rate is often accompanied by a loss-of-function process (pseudogeneization) for many copies through, for instance, mutations like nucleotide substitutions, which may introduce premature stop codons or indels, in turn generating frameshifts. This can lead to the presence of several gene copies sharing high sequence similarity even though some of them may contain nonsense mutations, frameshifts or be truncated. The annotation of gene copies harboring nonsense mutations is problematic compared with the initial unaltered copy. Some pipelines will be able to detect the entire coding phase but will introduce false introns to sidestep stop codons or frameshifts in order to retrieve a putatively translatable CDS (Figure 3b). Indeed, we noticed that the MSU automatic annotation tended to sidestep nonsense and frameshift mutations by introducing short introns. Such errors were observed previously by Meyers when re-annotating the *A. thaliana* NLR gene family (Meyers et al., 2003). Two arguments strengthen the assertion that such introns are false: (i) such introns are never found in more than one copy, whereas the intron positions are known to be well preserved between closely related copies; and (ii) sequence comparisons performed against recent paralogs (or orthologs from close relative species) have shown that the sequences of these wrongly annotated introns are always clearly homologous to coding sequences in other gene copies. Among the intron gain mechanisms, intronization (i.e. the process by which an exonic sequence is changed into an intron by mutation accumulation) is a complex process that is not yet very well documented or understood (Roy, 2016; Yenerall and Zhou, 2012). If it really occurs in genomes, it implies that a sufficiently long period of time must have passed for these mutations to occur and generate novel splicing sites. It is thus very unlikely that so many new introns arose in such a short period of time, as revealed by the low level of divergence between these genes and their paralog and/or ortholog counterparts. Other annotation pipelines, such as that of the IRGSP consortium, are more conservative in the sense that they give gene models with a more biologically meaningful expected structure, e.g. truncated proteins, in accordance with the presence of the first premature stop codons (either in-frame or caused by a frameshift; Figure 3b). This conservative choice could likely explain why we found more sequences classified as LRR-RLP from the IRGSP annotation than from the two other annotations (Table 1). Indeed, any LRR-RLK with a premature stop codon somewhere before the kinase domain would be considered as LRR-RLP (Figure 1). The annotation inconsistencies that we pinpointed here were observed for LRR-CR genes and did not question the overall quality of the three available rice genome annotations. They highlighted the limits of automatic annotation pipelines to annotate such complex multigene families and the consequences that given pipeline decision rules may have when drawing evolutionary conclusions.

The comparisons of the different gene models proposed by the three Nipponbare annotations led us to undertake a manual curation of the LRR-CR gene family. We are not the first to get involved in this painstaking but necessary work. Several high-quality studies have been based on re-annotated data, particularly in *A. thaliana* (Meyers et al., 2003; Van de Weyer et al., 2019). Expert annotations could also contain errors, of course, but expert curation limits their number. Opting exclusively for automated annotation should be avoided or otherwise operators should be aware that these annotations may contain errors induced by gene family specificities. These biases must be known and understood to avoid drawing misleading conclusions.

The expert Nipponbare rice genome contains more than 1000 LRR-CR loci, of which 30% have a non-canonical gene model

We curated LRR-CR loci in the reference Nipponbare rice genome by first comparing the three publicly available annotations at each locus: IRGSP, MSU and NCBI. Our aim was to retrieve LRR-CR genes in their entirety and account for the coding sequences as they probably stood before mutation accumulation. We obtained evidence that the sequence portions that we included in our gene models were not random genomic sequences but instead parts of the original gene CDS, as shown by the recovery of protein domains belonging to LRR-CR genes (kinase, NB-ARC, TM; Figure S4). Man et al. (2020) reported seven cases of missed domains through probable annotation errors in rice. We have also identified and corrected these seven genes and recovered the same domains. However, because our search for annotation errors was exhaustive, we recovered a higher number of missed domains.

When a gene had a nonsense mutation (in-frame stop codon or frameshift), an unexpected splicing site, or no terminal stop or start codon, we tagged it as non-canonical. This canonical versus non-canonical classification was based solely on features observed in gene models and did not imply any judgements on gene functionality. Genes tagged as non-canonical spanned a wide variety of cases, some of which could very likely not be translated into a functional protein while others may have had a function. As a first example, mutations inducing a premature stop codon could lead to a shorter protein that might sometimes perform the same function. Yet in many other cases shorter proteins might not perform the same function, if able to perform any function at all. Another example concerns the loss of the expected stop codon. When screening a sequence, a stop codon will eventually be encountered, but determining the functional consequences of this additional amino acid stretch would be impossible *in silico*. The same holds when the start codon is lost. Determining the criteria by which an alternative start codon (if any) may become the new start codon is a hazardous task. These few examples highlight the extent to which sorting out different functional scenarios is challenging. Moreover, mRNA molecules may play a regulating role, even if they cannot be translated as such, thereby justifying the need to annotate them. These reflections led us to voluntarily disregard such interpretations in our re-annotation process.

We observed that a third of the LRR-CR genes were non-canonical, but their proportion varied according to the gene subfamily (Table 2). A lower proportion of LRR-RLK genes were non-canonical (15%), compared with LRR-RLP (33%) and NLR (36%). The LRR-RLK subfamily could be divided into 15–20 subgroups based on phylogenetic study findings, and the duplication rate was shown to be quite variable according to the subgroup considered (Fischer et al., 2016; Tang et al., 2010). Some subgroups, the genes of which have been described as mostly involved in developmental processes, have had a more stable copy number over the course of angiosperm evolution (Fischer et al., 2016). These genes are less prone to duplication and thus are less likely to generate copies accumulating nonsense mutations, thereby lowering the proportion of non-canonical genes when the entire LRR-RLK subfamily is considered. The higher proportion of non-canonical genes obtained for NLR and LRR-RLP suggests that these subfamilies generally have higher birth and death rates. A quarter of the LRR-CR genes required manual curation and were non-canonical (representing 83.5% of the curated loci and 89.5% of the non-canonical loci; Table S2). In fact, manual curation was conducted mainly when none of the three annotations gave satisfactory gene models (such as the example presented in Figure 3b). The high correlation between non-canonical and curated loci was thus likely caused by the presence of mutations introducing ambiguities, which are overcome to different extents by the three annotation pipelines. A step forward would be to improve the annotation tools so that they deal differently with these nonsense mutations, e.g. including them in the sequences and indicating their presence without sidestepping them. To this end, annotation tools would have to predict non-canonical structures and tag them accordingly. To process such complex data, machine learning approaches are very promising (Mahood et al., 2020), but this implies having a significant learning corpus that has yet to be built.

We stress that this categorization of loci into canonical or non-canonical models could be impacted by the genomic sequence quality. Errors in the reference genome sequence could introduce errors in the gene models. In our curated data set, 27 non-canonical genes were tagged in NCBI data as harboring a difference between the RefSeq transcript sequence or protein and the Nipponbare reference sequence. The mutations jeopardizing the expected gene structure corresponded exactly to the positions where inconsistencies had been highlighted between the genomic and the RefSeq data. In order to appreciate the impact of errors when genes were categorized as non-canonical, we checked 65 of them in Nipponbare using both expression data and genome resequencing data. Only three probable errors were detected (one containing an 'N'), and four could not be validated. Moreover, among the 27 genes for which NCBI reported a potential error in the genomic sequence, 25 actually contained the identified nonsense mutation. Redundancies in LRR-CR gene sequences can give rise to ambiguities during both genome sequence assembly and expression data mapping, thus leading to errors (Torresen et al., 2019). Access to more specific re-sequencing data will resolve those potential inconsistencies. In the current state of the data, the reference genome errors identified concern less than 1% of the non-canonical genes.

LRR-CR repertoire in Kitaake, and comparison with Nipponbare

We propose a modular strategy to transfer our manually curated annotations to other rice genomes. We applied this strategy to annotate LRR-CR genes from the genome of the Kitaake cultivar, which also belongs to the *O. sativa* japonica subspecies. A comparison of the Nipponbare and Kitaake LRR-CR repertoires revealed an equivalent number of loci. The distributions of LRR-CR loci per gene subfamily, chromosome and category (canonical or not) were also consistent between these two cultivars (Figure 5).

In the Nipponbare genome, eight new LRR-CRs (one LRR-RLK, three LRR-RLPs and four UCs) were identified. These genes had not been previously annotated in any of the three publicly available annotations. In Kitaake, the same strategy enabled us to identify 114 new LRR-CR genes (48 of which were canonical). The higher number of unannotated LRR-CR genes in the Kitaake genome compared with the Nipponbare genome (114 versus eight) suggested that annotation inaccuracies had a greater impact on recently sequenced genomes that have not benefited from as much annotation investment as reference genomes.

A comparison of the LRR motif number for all allelic pairs between Nipponbare and Kitaake revealed a much greater difference in LRR number between alleles for publicly available annotations (ranging from 2.68 to 3.58), in comparison with our manually curated annotations (0.58), when the three subfamilies were all considered (Figure 6 and Figure S5). When publicly available annotations are considered, some rare allelic pairs harboring a very different number of LRRs may be truly different functional alleles. For instance, between two LRR-RLP alleles, one may contain a premature stop codon leading to the loss of a few motifs, but it may still have a biological function. It is important to identify such a pair. In our expert annotation alleles may share an identical number of LRRs, but such allelic pairs would be clearly identifiable because one of the alleles would be tagged as non-canonical whereas the other would be tagged as canonical. Moreover, in non-canonical alleles, the causal mutation, its position and impact on the gene (i.e. frameshift or premature stop codon) could be identified.

The difference in LRR motif number observed between allele pairs was greater for NLR than for the other subfamilies (Figure 6). This might be explained by the fact that NLR motifs are more variable and hence harder to detect (Ng et al., 2011), which could lead to apparent variations in the number of motifs in the two alleles. LRR motifs that have been found in NLRs differed from the common 'plant-specific' LRR consensus sequence, and were more irregular in terms of both length and residue conservation (Kajava, 1998; Kuang et al., 2004; Matsushima and Miyashita, 2012; Sela et al., 2012). Although we enhanced the LRR detection accuracy through the development of a new LRR HMM profile for NLR, it is still not exhaustive. This also suggests that the number of LRR motifs varies more in NLR than in other LRR-CR subfamilies.

High sequence similarity was observed between Nipponbare and Kitaake alleles (98.9% identity for cDNA) (Table S3), which was consistent with previous comparisons (Jain et al., 2019). However, some allelic pairs showed a lower identity level with a more ancient coalescent history between the two genomes. This heterogeneity may have been the consequence of the breeding programs from which these varieties were derived. The breeding process involves crosses with more or less closely related genotypes, sometimes from different subspecies, and may generate mosaic genomes (Santos et al., 2019). No allelic pairs were found for 106 genes: i.e. 48 were specific to Nipponbare and 58 were specific to Kitaake. A majority of those genes were located in clusters on chromosomes 2, 11 and 12, which have already been described as containing a large number of LRR-CRs (Mizuno et al., 2020; Zhou et al., 2004) (Figure 7). Some clusters have also been shown to be less conserved (Mizuno et al., 2020). More than half of these genes were classified as canonical.

The methods we developed allowed us to undertake an exhaustive comparison of the LRR-CR repertoire between Nipponbare and Kitaake. Allelic pairs, including those hosting nonsense mutations in either or both genotypes, were described (Data S4). Genotype-specific genes were also identified and localized, again along with information related to the potential presence of nonsense mutations (Figure 7). These results were achieved through a combination of an expert annotation and its transfer to a second genotype for which a high quality *de novo* genome assembly was available. Validation of the LRR-CR annotations of Kitaake was not very time consuming compared with the initial work in Nipponbare, where each gene was investigated individually. Our study highlighted that investment in a combination of technologies would guarantee high-quality assemblies and annotations, especially when the discovery of allelic diversity is targeted (Zhou et al., 2020).

The tools and curated data sets that we generated in this study are available from: https://rice-genome-hub.southgreen.fr/content/geloc (data) and https://github.com/cgottin/LRRprofiler (tools). Note that we focused on developing a website where stop codons and frameshifts are easily identified. We believe that evolutionary studies and allele discovery initiatives for LRR-CRs would be more accurate and reliable when using our manually curated comprehensive annotations for these genes. Moreover, we feel that this comprehensive annotation approach should be widely adopted by the community in the light of the major potential benefits it provides.

EXPERIMENTAL PROCEDURES

Genomes and annotation files

Reference genomic sequences of Nipponbare (Kawahara et al., 2013) and Kitaake (Jain et al., 2019) *O. sativa* ssp. *japonica* cultivars were downloaded from the Rice Annotation Project Database (RAP-DB) website (https://rapdb.dna.affrc.go.jp) and the Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html). The general feature format (GFF) and fasta files with coding DNA sequences (CDSs) and protein sequences for Nipponbare were downloaded for three different annotation projects: (i) the MSU 7.0 annotation was downloaded from the Rice Genome Annotation Project FTP server (http://rice.plantbiology.msu.edu); (ii) the IRGSP annotation files were downloaded from the RAP-DB website (https://rapdb.dna.affrc.go.jp); and (iii) the NCBI annotation (release 102) annotated by the NCBI Eukaryotic Genome Annotation Pipeline was downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov). The IRGSP annotation consists of two gene sets ('genes supported by FL-cDNAs, ESTs or proteins' and 'computationally predicted genes') that were concatenated for the analyses (Sakai et al., 2013).

LRRPROFILER implementation

The LRRPROFILER pipeline was implemented in two steps (Figure S1). The first step involved the iterative refinement of LRR HMM profiles specific to a gene subfamily (LRR-RLK or NLR) and proteome (Figure S1; inspired by Ng et al., 2011). Only LRR-RLK and NLR were considered for profile refinement because they contain a specific domain (i.e. kinase and NB-ARC domains, respectively), thereby allowing the clear identification of the subfamily to which they belong. A set of candidate protein sequences was identified from a given proteome to refine the specific LRR profiles. This set was composed of either LRR-RLKs identified with iTAK (Zheng et al., 2016) or NLRs identified with the PF00931 Pfam NB-ARC profile. A first round of LRR motif detection was performed in either of the candidate protein sets using HMMSEARCH (HMMER; Eddy, 2011) with the SM00370 LRR profile from the SMART database. Motifs of 20–26 amino acids in length were extracted, aligned with MAFFT (Katoh and Standley, 2013) with default parameters and a new profile was built from the alignment using HMMBUILD (HMMER; Eddy, 2011). This process was repeated using the HMM LRR profile built at the previous iteration to search again for LRR motifs in the considered protein candidate set. At each iteration, the sum of the amino acid lengths of the detected LRR motifs was calculated. The process stopped when three iterations (not necessarily consecutive) resulted in a decrease of the statistics. Finally, the process retrieved the HMM LRR profile identifying the maximum number of LRR motifs in the candidate protein set.

The second step of the LRRPROFILER pipeline consisted of the identification of LRR-CR proteins present in a given proteome, the annotation of their functional domains as well as their classification into a gene subfamily: LRR-RLK, LRR-RLP, NLR or UC (Figure S1). Six publicly available LRR HMM profiles from the SMART database, i.e. SM00364 (LRR_BAC), SM00365 (LRR_SD22), SM00367 (LRR_CC), SM00368 (LRR_RI), SM00369 (LRR_TYP) and SM00370 (LRR), in addition to the newly built LRR profiles obtained in the first step were used to detect LRR motifs in the complete proteome under consideration using HMMSEARCH. An annotation of the LRR domains, containing the start and end positions of each LRR motif, was part of the output. The annotation of each protein was then supplemented using publicly available profiles for other functional domains: TIR (PF01582), TIR_2 (PF13676), Malectin (PF11721), Malectin-like (PF12819), RPW8 (PF05659), Cys-Pairs (Dievart and Clark, 2003; Dufayard et al., 2017), F-box (PF00646) and FBD (PF08387). NB-ARC and kinase domain annotations were retrieved from the first step, whereas transmembrane domains (TMs) were detected with TMHMM 2.0C, with default parameters (Sonnhammer et al., 1998). The subfamily assignment of each identified LRR-containing protein was deduced from its domain structure. Proteins were classified into the LRR-RLK subfamily if they contained at least one LRR motif and a kinase domain, and sometimes other domains such as the malectin, malectin-like, Cys-pair and TM domains. Proteins were classified in the NLR subfamily if they included an NB-ARC domain and at least one LRR motif, sometimes with a TIR or an RPW8 domain. The LRR-RLP subfamily included proteins with LRRs plus a TM, malectin, malectin-like and/or Cys-pair, or LRR-only structures when at least 13 plant-specific LRR repeats were detected. Proteins containing an F-box or an FBD domain in addition to LRRs were classified as F-box-LRR. All other LRR-containing proteins were ranked in the UC group, and for these we performed a

BLASTp search with default parameters against the other gene sets (LRR-RLP, LRR-RLK, NLR and F-box) to estimate their probable membership of one of these gene subfamilies. F-box proteins were removed from our data sets and not considered further in the analyses. We ended up with four gene sets: LRR-RLP, LRR-RLK, NLR and UC.

At the end of the construction phase, the complete LRRPROFILER pipeline was tested on the manually reviewed *A. thaliana* protein data set downloaded from the Swiss-Prot section (https://www.uniprot.org) (Data S1; Figure S2; Methods S1; Table S1) (Boutet et al., 2007) of the UniProt databank (The UniProt Consortium, 2019). This set was composed of 15 818 sequences. Domain and repeat information was also extracted from the database, in particular the number of LRR motifs per sequence and the gene subfamily to which it belonged (LRR-RLP, LRR-RLK, NLR, etc.).

Rice transcription factor data set

Transcription factor genes (TFs) were identified in the proteome predicted from the three publicly available annotations of the Nipponbare rice reference genome using ITAK (Zheng et al., 2016). Nine subfamilies were considered: C2H2, FAR1, MYB-related, WRKY, NAC, AP2/ERF-ERF, bHLH, bZIP and MYB.

Annotation transfer from Nipponbare to Kitaake

The first phase consisted of locating regions of interest in the Kitaake genome, i.e. regions homologous to Nipponbare LRR-CR loci (Figure 4a). Nipponbare LRR-CR protein sequences from our expert annotations were aligned with the Kitaake genome using tBLASTn (Altschul et al., 1990). Only high scoring pairs (HSPs) with more than 65% identity with Nipponbare LRR-CR protein fragments and spanning at least 50% of the Nipponbare query protein were retained. To define coherent candidate regions in Kitaake, HSPs from the same Nipponbare query protein had to be located less than 5000 bp apart, except when the Nipponbare homologous gene queried had a longer intron. In that case, the Nipponbare intron length plus 500 bp was used as the upper bound for the distance separating Kitaake HSPs. Multiple regions of interest could be found for a single Nipponbare protein. This allowed us to annotate genes duplicated in the Kitaake genome even if a single gene copy was present in the Nipponbare genome. In a second phase, gene model determination was attempted in three consecutive steps for each region of interest (Figure 4b). Only regions that could not be successfully annotated at a given step passed to the next step. In the first step, the Nipponbare query exons are mapped to the target Kitaake region of interest with BLASTn. A gene model was then reconstructed based on ordered HSPs. The gene model reconstruction quality was checked by comparing the predicted protein with that of Nipponbare using BLASTp. The gene model was retained if all expected exons were present and the Kitaake protein sequence had more than 90% identity with the Nipponbare protein sequence. Otherwise, the annotation of this region was delegated to the second step. In the second step, the EXONERATE cdna2genome model (Slater and Birney, 2005) was run independently for every remaining query/target pair. The EXONERATE output GFF file was parsed to construct the target gene model and to document putative frameshift positions. Again, the Kitaake predicted protein was compared with the Nipponbare query sequence with BLASTp and retained if the coverage and identity were above 90 and 75%, respectively. Otherwise, the annotation of this target region was delegated to the third step. In the third step, the remaining loci were reconstructed with the EXONERATE protein2genome model. This model is better at finding the correct reading frames when the target and model loci are more divergent, but it fails to correctly annotate type-1 and -2 splicing sites (intron/exon junction falling inside a codon). This problem arises because it uses the same reading frame to translate the whole genomic sequence (the six reading frames are tested, but each resulting translation just uses one of them). To overcome this issue, intron junctions are then corrected with a PYTHON script that looks for canonical splicing sites in a range of two nucleotides before and after the current junctions. Finally, gene models highly divergent from the Nipponbare query sequence, with multiple premature stop codons or without start or terminal stop codons, and overlapping frameshifts are tagged to be checked manually.

Identification of alleles between Nipponbare and Kitaake

We used SYNMAP (Lyons and Freeling, 2008) to identify LRR-CR allelic pairs, i.e. genes with the exact same chromosomic position in Nipponbare and Kitaake. SYNMAP was developed to identify orthologous genes between different species based on microcolinearity conservation, and it identifies blocks of genes of conserved order and position. It retrieves a list of relationships between genic repertoires of two genomes. We identified alleles by first selecting genes for which SYNMAP found a reciprocal relationship, i.e. a relationship found in both Nipponbare–Kitaake and Kitaake–Nipponbare comparisons. Genes for which allelic relationships could not be unambiguously resolved by SYNMAP were manually resolved, when possible, using VISTA (Mayor et al., 2000) and ARTEMIS (Carver et al., 2012).

ACKNOWLEDGMENTS

This work was partly supported by the CGIAR Research Program on Rice (CRP RICE) (to AD) and by a PhD fellowship from Institut Agro and CIRAD (to CG). Computational resources were provided by the South Green bioinformatics platform. The authors would like to thank Dr A. Cenci for a critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

CG, NC, AD, CP and VR designed the research. CG, NC and AD performed the research. CG, NC, AD, VR, MS and GD contributed to new analytic and computational tools. CG, NC and AD analyzed the data. CG, NC, AD and VR wrote the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest associated with this work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Schematic representation of the LRRPROFILER pipeline.

Figure S2. Comparison of expected and predicted LRR motifs in protein sequences from the Swiss-Prot *Arabidopsis thaliana* data set using publicly available and refined HMM profiles.

Figure S3. Comparison of predicted peptide lengths between Nipponbare publicly available annotations (IRGSP, MSU and NCBI) for LRR-CR and TF loci.

Figure S4. Number of domains and motifs identified with LRRPROFILER for the Nipponbare proteomes predicted by publicly available and manually curated annotations.

Figure S5. LRR motif number conservation between Nipponbare and Kitaake LRR-CR loci, depending on the annotation compared.

Table S1. Performance of publicly available and refined LRR HMM profiles in the Swiss-Prot *Arabidopsis thaliana* data set.

Table S2. Contingency table of canonical/non-canonical and modified/not modified LRR-CR loci from the Nipponbare manually curated annotation.

Table S3. Percentage of cDNA identity between Nipponbare and Kitaake alleles according to gene subfamilies and categories.

Methods S1. Validation of the LRRPROFILER pipeline.

Data S1. LRRPROFILER results in the Swiss-Prot *Arabidopsis thaliana* data set.

Data S2. LRR-CR loci from the Nipponbare rice reference genome.

Data S3. LRR-CR loci from the rice KitaakeX genome.

Data S4. Allelic relationship and cDNA identity between Nipponbare and Kitaake LRR-CR loci.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at the detailed link as follows: https://doi.org/10.5281/zenodo.5110015

DATA AVAILABILITY STATEMENT

All of the data files (gff and fasta files) are available from the dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc) and from the open data repository Zenodo (https://doi.org/10.5281/zenodo.5110015).

A new identifier was allocated to each LRR-CR gene unraveled by this procedure. These identifiers use the <OSJnip_ChrXX_00000000> or <OSJkit_ChrXX_00000000> pattern for Nipponbare and Kitaake loci, respectively, with XX being the chromosome number followed by the start codon position of the coding sequence (CDS) on the chromosome (Data S2 and S3).

According to the Multiple Alignment of Coding Sequences (MACSE) convention (Ranwez et al., 2018; Ranwez et al., 2011), indels causing frameshift mutations have been pinpointed by the presence of one or two '!' characters in the nucleotide sequences of non-canonical genes and are available in an additional specific data set.

REFERENCES

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S. et al. (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Andersen, E.J., Nepal, M.P., Purintun, J.M., Nelson, D., Mermigka, G. & Sarris, P.F. (2020) Wheat disease resistance genes and their diversification through integrated domain fusions. *Frontiers in Genetics*, **11**, 898.

Bailey-Serres, J., Parker, J.E., Ainsworth, E.A., Oldroyd, G.E.D. & Schroeder, J.I. (2019) Genetic strategies for improving crop yields. *Nature*, **575**, 109–118.

Bayer, P.E., Edwards, D. & Batley, J. (2018) Bias in resistance gene prediction due to repeat masking. *Nature Plants*, **4**, 762–765.

Bella, J., Hindle, K.L., McEwan, P.A. & Lovell, S.C. (2008) The leucine-rich repeat structure. *Cellular and Molecular Life Sciences*, **65**, 2307–2333.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. (2007) UniProtKB/Swiss-Prot. *Methods in Molecular Biology*, **406**, 89–112.

Boutrot, F. & Zipfel, C. (2017) Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annual review of Phytopathology*, **55**, 257–286.

Burdett, H., Bentham, A.R., Williams, S.J., Dodds, P.N., Anderson, P.A., Banfield, M.J. et al. (2019) The plant "Resistosome": structural Insights into Immune Signaling. *Cell Host & Microbe*, **26**, 193–201.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J. & McQuillan, J.A. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.

Couto, D. & Zipfel, C. (2016) Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology*, **16**, 537–552.

Dievart, A. & Clark, S.E. (2003) Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Current Opinion in Plant Biology*, **6**, 507–516.

Dufayard, J.F., Bettembourg, M., Fischer, I., Droc, G., Guiderdoni, E., Perin, C. et al. (2017) New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms. *Frontiers in Plant Science*, **8**, 381.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

FAO. (2018) The future of food and agriculture—Alternative pathways to 2050 Rome.

Fawal, N., Li, Q., Mathe, C. & Dunand, C. (2014) Automatic multigenic family annotation: risks and solutions. *Trends in Genetics*, **30**, 323–325.

Fischer, I., Dievart, A., Droc, G., Dufayard, J.F. & Chantret, N. (2016) Evolutionary dynamics of the leucine-rich repeat receptor-like kinase (LRR-RLK) subfamily in angiosperms. *Plant Physiology*, **170**, 1595–1610.

Fritz-Laylin, L.K., Krishnamurthy, N., Tor, M., Sjolander, K.V. & Jones, J.D. (2005) Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis. *Plant Physiology*, **138**, 611–623.

Furumizu, C. & Sawa, S. (2021) Insight into early diversification of leucine-rich repeat receptor-like kinases provided by the sequenced moss and hornwort genomes. *Plant Molecular Biology*. Online ahead of print. https://doi.org/10.1007/s11103-020-01100-0

Han, G.Z. (2019) Origin and evolution of the plant immune system. *New Phytologist*, **222**, 70–83.

Hosseini, S., Schmidt, E.D.L. & Bakker, F.T. (2020) Leucine-rich repeat receptor-like kinase II phylogenetics reveals five main clades throughout the plant kingdom. *The Plant Journal*, **103**, 547–560.

Hu, J., Chang, X., Zou, L., Tang, W. & Wu, W. (2018) Identification and fine mapping of Bph33, a new brown planthopper resistance gene in rice (Oryza sativa L.). *Rice*, **11**, 55.

Hwang, S.G., Kim, D.S. & Jang, C.S. (2011) Comparative analysis of evolutionary dynamics of genes encoding leucine-rich repeat receptor-like kinase between rice and Arabidopsis. *Genetica*, **139**, 1023–1032.

Innan, H. & Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, **11**, 97–108.

Jain, R., Jenkins, J., Shu, S., Chern, M., Martin, J.A., Copetti, D. et al. (2019) Genome sequence of the model rice variety KitaakeX. *BMC Genomics*, **20**, 905.

Jones, D.A. & Jones, J.D.G. (1997) The role of leucine-rich repeat proteins in plant defences. In *Advances in Botanical Research* (Andrews, J.H., Tommerup, I.C. & Callow, J.A., eds). Academic Press, pp. 89–167.

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G.J. et al. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *The Plant Journal*, **76**, 530–544.

Kajava, A.V. (1998) Structural diversity of leucine-rich repeat proteins. *Journal of Molecular Biology*, **277**, 519–527.

Kajava, A.V. (2012) Tandem repeats in proteins: from sequence to structure. *Journal of Structural Biology*, **179**, 279–288.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S. et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E. & Michelmore, R.W. (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *The Plant Cell*, **16**, 2870–2894.

Lai, X., Chahtane, H., Martin-Arevalillo, R., Zubieta, C. & Parcy, F. (2020) Contrasted evolutionary trajectories of plant transcription factors. *Current Opinion in Plant Biology*, **54**, 101–107.

Lee, H.Y., Mang, H., Choi, E., Seo, Y.E., Kim, M.S., Oh, S. et al. (2021) Genome-wide functional analysis of hot pepper immune receptors reveals an autonomous NLR clade in seed plants. *New Phytologist*, **229**, 532–547.

Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends in Genetics*, **20**, 116–122.

Li, J., Ding, J., Zhang, W., Zhang, Y., Tang, P., Chen, J.Q. et al. (2010) Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. *Molecular Genetics and Genomics*, **283**, 427–438.

Li, L.-Y., Wang, L., Jing, J.-X., Li, Z.-Q., Lin, F., Huang, L.-F. et al. (2007) The Pikm gene, conferring stable resistance to isolates of Magnaporthe oryzae, was finely mapped in a crossover-cold region on rice chromosome 11. *Molecular Breeding*, **20**, 179–188.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. & You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, **17**, 852.

Lyons, E. & Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, **53**, 661–673.

Mahood, E.H., Kruse, L.H. & Moghe, G.D. (2020) Machine learning: a powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, **8**, e11376.

Man, J., Gallagher, J.P. & Bartlett, M. (2020) Structural evolution drives diversification of the large LRR-RLK gene family. *New Phytologist*, **226**, 1492–1505.

Martin, E.C., Sukarta, O.C.A., Spiridon, L., Grigore, L.G., Constantinescu, V., Tacutu, R. et al. (2020) LRRpredictor-A new LRR motif detection method for irregular motifs of plant NLR proteins using an ensemble of classifiers. *Genes*, **11**, 286.

Matsushima, N. & Miyashita, H. (2012) Leucine-rich repeat (LRR) domains containing intervening motifs in plants. *Biomolecules*, **2**, 288–311.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A. et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.

McDowell, J.M. & Simon, S.A. (2006) Recent insights into R gene evolution. *Molecular Plant Pathology*, **7**, 437–448.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *The Plant Cell*, **15**, 809–834.

Michelmore, R.W. & Meyers, B.C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research*, **8**, 1113–1130.

Mizuno, H., Katagiri, S., Kanamori, H., Mukai, Y., Sasaki, T., Matsumoto, T. et al. (2020) Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. *Scientific Reports*, **10**, 872.

Nei, M. & Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, **39**, 121–152.

Ng, A.C., Eisenberg, J.M., Heath, R.J., Huett, A., Robinson, C.M., Nau, G.J. et al. (2011) Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proceedings of the National Academy of Sciences U S A*, **108**(Suppl 1), 4631–4638.

Okuyama, Y., Kanzaki, H., Abe, A., Yoshida, K., Tamiru, M., Saitoh, H. et al. (2011) A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *The Plant Journal*, **66**, 467–479.

Prigozhin, D.M. & Krasileva, K.V. (2021) Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. *The Plant Cell*, **33**, 998–1015.

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N. & Delsuc, F. (2018) MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, **35**, 2582–2584.

Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J. (2011) MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One*, **6**, e22594.

Rice Full-Length cDNA Consortium, National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team; Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K. et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from Japonica rice. *Science*, **301**, 376–379.

Richter, T.E. & Ronald, P.C. (2000) The evolution of disease resistance genes. *Plant Molecular Biology*, **42**, 195–204.

Roy, S.W. (2016) How common is parallel intron gain? Rapid evolution versus independent creation in recently created introns in daphnia. *Molecular Biology and Evolution*, **33**, 1902–1906.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y. et al. (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant and Cell Physiology*, **54**, e6.

Sakamoto, K., Tada, Y., Yokozeki, Y., Akagi, H., Hayashi, N., Fujimura, T. et al. (1999) Chemical induction of disease resistance in rice is correlated with the expression of a gene encoding a nucleotide binding site and leucine-rich repeats. *Plant Molecular Biology*, **40**, 847–855.

Santos, J.D., Chebotarov, D., McNally, K.L., Bartholome, J., Droc, G., Billot, C. et al. (2019) Fine scale genomic signals of admixture and alien introgression among Asian rice landraces. *Genome Biology and Evolution*, **11**, 1358–1373.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N. & Nelson, A. (2019) The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, **3**, 430–439.

Sekhwal, M.K., Li, P., Lam, I., Wang, X., Cloutier, S. & You, F.M. (2015) Disease resistance gene analogs (RGAs) in plants. *International Journal of Molecular Sciences*, **16**, 19248–19290.

Sela, H., Spiridon, L.N., Petrescu, A.J., Akerman, M., Mandel-Gutfreund, Y., Nevo, E. et al. (2012) Ancient diversity of splicing motifs and protein surfaces in the wild emmer wheat (Triticum dicoccoides) LR10 coiled coil (CC) and leucine-rich repeat (LRR) domains. *Molecular Plant Pathology*, **13**, 276–287.

Shao, Z.Q., Wang, B. & Chen, J.Q. (2016) Tracking ancestral lineages and recent expansions of NBS-LRR genes in angiosperms. *Plant Signaling & Behavior*, **11**, e1197470.

Shiu, S.H. & Bleecker, A.B. (2001a) Plant receptor-like kinase gene family: diversity, function, and signaling. *Science STKE*, **2001**, re22.

Shiu, S.H. & Bleecker, A.B. (2001b) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. *Proceedings of the National Academy of Sciences U S A*, **98**, 10763–10768.

Slater, G.S. & Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T. et al. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, **270**, 1804–1806.

Sonnhammer, E.L., von Heijne, G. & Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **6**, 175–182.

Stanke, M. & Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl 2), ii215–ii225.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C. et al. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics*, **50**, 285–296.

Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.J., Yu, G. et al. (2020) The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiology*, **183**, 468–482.

Sun, X. & Wang, G.L. (2011) Genome-wide identification, characterization and phylogenetic analysis of the rice LRR-kinases. *PLoS One*, **6**, e16079.

Tamborski, J. & Krasileva, K.V. (2020) Evolution of plant NLRs: from natural history to precise modifications. *Annual Review of Plant Biology*, **71**, 355–378.

Tang, P., Zhang, Y., Sun, X., Tian, D., Yang, S. & Ding, J. (2010) Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species. *Plant Science*, **179**, 399–406.

The UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506–D515.

Thilmony, R., Guttman, M., Thomson, J.G. & Blechl, A.E. (2009) The LP2 leucine-rich repeat receptor kinase gene promoter directs organ-specific, light-responsive expression in transgenic rice. *Plant Biotechnology Journal*, **7**, 867–882.

Torresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P. et al. (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, **47**, 10994–11006.

Van de Weyer, A.L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K. et al. (2019) A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. *Cell*, **178**, 1260–1272.e14

van der Burgh, A.M. & Joosten, M. (2019) Plant immunity: thinking outside and inside the box. *Trends in Plant Science*, **24**, 587–601.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wilming, L. & Harrow, J. (2009) Gene annotation methods. In *Bioinformatics* (Edwards, D., Stajich, J. & Hansen, D., eds). New York, NY: Springer.

Xiong, Y., Han, Z. & Chai, J. (2020) Resistosome and inflammasome: platforms mediating innate immunity. *Current Opinion in Plant Biology*, **56**, 47–55.

Yao, W., Li, G., Yu, Y. & Ouyang, Y. (2018) funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *Gigascience*, **7**, 1–9.

Yenerall, P. & Zhou, L. (2012) Identifying the mechanisms of intron gain: progress and trends. *Biology Direct*, **7**, 29.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R. et al. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Research*, **31**, 229–233.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P. et al. (2016) iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant*, **9**, 1667–1670.

Zhou, T., Wang, Y., Chen, J.Q., Araki, H., Jing, Z., Jiang, K. et al. (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Molecular Genetics and Genomics*, **271**, 402–415.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S. et al. (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data*, **7**, 113.

RESEARCH ARTICLE

Domestication reduces alternative splicing expression variations in sorghum

Vincent Ranwez[1], Audrey Serra[1], David Pot[2], Nathalie Chantret[3]*

1 Montpellier SupAgro, UMR AGAP, Montpellier, France, **2** CIRAD, UMR AGAP, Montpellier, France, **3** INRA, UMR AGAP, Montpellier, France

* nathalie.chantret@inra.fr

Citation: Ranwez V, Serra A, Pot D, Chantret N (2017) Domestication reduces alternative splicing expression variations in sorghum. PLoS ONE 12 (9): e0183454. https://doi.org/10.1371/journal.pone.0183454

Editor: Manoj Prasad, National Institute of Plant Genome Research, INDIA

Received: April 19, 2017

Accepted: August 6, 2017

Published: September 8, 2017

Data Availability Statement: All RNAseq raw data are freely available on the NCBI RSA database (Sequence Read Archive) (cultivated: SAMN05277472 to SAMN05277481; wild: SAMN06052464 to SAMN06052472 and SAMN07313361). Other data (1) can be freely downloaded from http://bioweb.supagro.inra.fr/AlternativeSplicing/ and (2) are within the paper and its Supporting Information files (S1_Table and S1_File).

Funding: This study was conducted in the framework of ARCAD (http://www.arcad-project.

Abstract

Domestication is known to strongly reduce genomic diversity through population bottlenecks. The resulting loss of polymorphism has been thoroughly documented in numerous cultivated species. Here we investigate the impact of domestication on the diversity of alternative transcript expressions using RNAseq data obtained on cultivated and wild sorghum accessions (ten accessions for each pool). In that aim, we focus on genes expressing two isoforms in sorghum and estimate the ratio between expression levels of those isoforms in each accession. Noticeably, for a given gene, one isoform can either be overexpressed or underexpressed in some wild accessions, whereas in the cultivated accessions, the balance between the two isoforms of the same gene appears to be much more homogenous. Indeed, we observe in sorghum significantly more variation in isoform expression balance among wild accessions than among domesticated accessions. The possibility exists that the loss of nucleotide diversity due to domestication could affect regulatory elements, controlling transcription or degradation of these isoforms. Impact on the isoform expression balance is discussed. As far as we know, this is the first time that the impact of domestication on transcript isoform balance has been studied at the genomic scale. This could pave the way towards the identification of key domestication genes with finely tuned isoform expressions in domesticated accessions while being highly variable in their wild relatives.

Introduction

Alternative splicing (AS) is the mechanism by which two or more processed mRNA isoforms result from the maturation of the same primary transcribed precursor mRNA molecule (pre-mRNA) [1]. One of the main steps of the pre-mRNA maturation is the splicing process, during which introns are removed from the pre-mRNA molecule, orchestrated by a whole array of *trans*-acting regulator proteins as well as *cis*-acting elements within the pre-mRNA itself. Occurring in all eukaryotes, AS has been extensively described and studied in humans [2] and other animals [3]. Through increasing diversity and complexity of transcriptomes, AS has two major outcomes: proteome diversification and regulation of gene expression. AS was suggested to be one of the possible origins of the large phenotypic differences among species which otherwise share a similar repertoire of protein-coding genes, as vertebrates do, for example [3].

org), a project funded by Agropolis Fondation under the reference ID ARCAD 0900-001.

Competing interests: The authors have declared that no competing interests exist.

AS is recognised to be a "pivotal step between transcription and translation" [4]. It has been described as varying according to organ, according to developmental stages and even according to cell type [5]. AS complex regulation is the guarantee of a consistent development for a given organism [6], several AS misregulations have been identified as causing diseases [7, 8]. Its role has been increasingly pointed out as a key factor of regulation in animals. The question of its prevalence in plants was much slower to emerge [9]. At the beginning of the last decade, AS started to be investigated in model species at the scale of the genome. The proportion of genes described as affected by AS has increased following the progress in sequencing technologies, to reach values of 48% and 61% of the intron-containing genes for recent estimations in rice [10] and *Arabidopsis thaliana* [11] respectively. Since RNAseq data is getting easier and cheaper to use, and bioinformatic tools are now available to process data and predict AS events (e.g. [12]) AS is now described, at the genomic scale, for many more species: *Brachypodium distachyon* [13], *Vitis vinifera* [14], *Hordeum vulgare* [15], tomato [16] and sorghum [17] to cite only a few of them. Some comparative analyses of AS have now started to be carried out on several species [18].

Regardless of the studied organism, the proportion of mRNA isoforms identified that are actually translated into functional proteins is not precisely known [19] and AS impact on plant proteome diversification is still being debated [20]. However, owing to the diversity and complexity of mRNA molecules AS generates, it is believed to play an essential role in the regulation of expression, and/or to affect translation probability, via the nonsense-mediated decay (NMD). NMD is a process during which alternatively spliced isoforms possessing a premature stop codon are degraded [21, 22]. Indeed AS induced regulation is very sensitive to environmental conditions. It has been shown that important changes in AS patterns occur in plants in response to environmental stresses (recently reviewed in [23, 24, 25]). A steady stream of new papers continuously brings additional examples of the role of AS in mechanisms involved in stress responses [26, 27]. Finally AS has been shown to play a role in plant immunity, through plant disease resistance genes (R-genes) AS (reviewed in [28]).

Although AS plays a key role in several biological processes, the question of its intraspecific variability has been raised only recently and only a few cases of plant intraspecific variability have been studied so far. In a recent study, Potenza et al. explored the AS landscape in ten grapevine cultivars [14]. They found that the AS isoforms are well conserved across individuals with up to 21% of them conserved across the 10 genotypes despite the fact that in most cases (~70%) one isoform is expressed at least ten times less strongly than the canonical forms. An open question remains concerning AS isoform repertoire variation among cultivars possibly due to variability in the splicing sites or possibly to the fine tuning of the spliceosome machinery (other regulatory elements, *cis* or *trans*), or both.

Up to now, how AS is finely tuned in a given individual, organ, or developmental step, is not known but the mere fact that AS varies according to genotypes and environmental changes [24] is a clue to its potential role in genetic adaptation. Consequently, one could wonder whether crop and animal domestication has a significant impact on the pattern of variability of AS.

All the traits making the crop different from its wild relative are grouped under the term of 'domestication syndrome'. In the case of plants, this includes changes in secondary metabolites, modifications of plant architecture, increases in fruit size, loss of seed dormancy and alteration of dispersion capacity, to cite only the main changes. However, it is quite variable according to species, and in particular, annual crops, such as sorghum, have been shown to exhibit significantly stronger domestication syndrome than perennial ones [29]. From a genetic standpoint, domestication is a combination of genetic drift effects caused by founder sampling (the strength of the resulting bottleneck varies according to species), and of selective effects caused by the deliberate selection of alleles for the advantage they confer for human uses [30]. One of the recurrent objectives is to identify the underlying genetic architecture of adaptation and ultimately the genes controlling physiological and morphological traits for which changes are observed between crops and their wild relatives [31]. The search for such genetic/phenotypic relationships is routinely done using Quantitative Trait Locus (QTL) mapping, genome wide association study (GWAS) or selection scan approaches, although the latest do not directly explore the statistical links between allelic and phenotypic diversity.

Finally, beyond the methods aiming to correlate genetic to phenotypic variations caused by domestication, recent studies have focused on intermediate steps lying between genetic and phenotype, gene expression, in particular. Expression of 18,242 genes was surveyed in maize and teosinte, its wild ancestor [32, 33]. Changes in expression levels were observed for 600 of them, but at the genome-wide scale, the coefficient of variation of expression among lines was not significantly different in maize and teosinte [33]. When considering the subset of 'candidate genes' located in regions that they identify as undergoing either domestication or posterior selection, they observed a reduced variation in expression levels in maize *versus* teosinte. This could suggest that *cis*-acting regulatory regions were affected by domestication [32]. In cotton, comparative gene expression showed a parallel up-regulation of several genes of the same gene family in independently domesticated cotton species [34]. In tomato, comparative transcriptomics revealed expression divergence between cultivated and wild accessions, and a correlation between network rewiring and light responsiveness in domesticated tomato [35]. In common bean a very clear decrease of gene expression variability (18%) was also detected in domesticated beans as compared to their wild counterparts [36]. Another strategy is to focus on the transcriptome of organs which underwent major morphological changes during domestication such as glumes in wheat [37] for which decreased expression levels of genes involved in cell walls, lignin, pectins and wax biosynthesis potentially contribute to the divergence of the glume's properties between wild and cultivated wheat. In cotton, it was shown that domestication affected the expression of many genes in fiber cells, with twice as many genes differentially expressed in fiber cell development in domesticated cotton versus wild [38]. This approach may help to understand the biological mechanisms underlying the complex links between genotype and phenotype, even if the causal mutation(s) controlling the difference of expression is (are) not identified. Additionally, as gene expression is an 'intermediate' trait, its analysis may help to identify genes that would have been missed through exclusive final phenotype variability analysis due to a lack of statistical power. Finally, a recent study identified a subset of genes expressing more isoforms in maize than in teosinte (wild relative of maize) but found no significant difference between their AS isoform repertoires (i.e. type of alternative splicing events: intron retention, alternative acceptor site and so on) [39]. However, whether domestication has impacted alternative splicing expression variability, and how, has not been described up to now.

In this paper, we study the impact of sorghum domestication on alternative splicing by identifying whether differential patterns of isoform expression are observed when comparing cultivated and wild compartments. Sorghum currently ranks fifth for grain production tonnage, providing staple food for 500 million people worldwide [40]. Its success is mainly due to its high level of drought tolerance and to its adaptation to a large spectrum of environmental conditions and uses. The recent release of its genome sequence [41], its phylogenetic proximity with several important C4 species (maize, switchgrass, sugarcane) and its low genome complexity contribute to its interest on a more fundamental level.

The *Sorghum bicolor* species includes three sub-species: ssp. *bicolor* (the domesticated form), ssp. *verticilliflorum* (the closest wild relative) and ssp. *drumondii* (the weedy form which corresponds to stable hybrids between the wild relatives and the cultivated types). The wild and domesticated pools are inter-fertile and intense gene flows occur (e.g. [42–45]). However a clear domestication syndrome is visible between the wild and cultivated pools. A key phenotypic difference between the cultivated and wild sorghum forms, controlled by the *SH1* gene [46], is that the cultivated type has large non shattering seeds whereas the wild type has small shattering seeds. Other traits corresponding to plant architecture (tillering), seed weight etc. are also highly divergent between these pools.

Concerning the mating system, the cultivated form does less outcrossing than the wild one, but even if selfing is predominant, outcrossing can reach up to 20% in some cultivated races such as the Guinea [47].

According to Hamblin [48], the domestication history of sorghum is complex and cannot be summarized by a single bottleneck event. Such a simple model simply does not fit their data and more complex scenario, e.g. including multiple domestications or introgression from wild congeners, have to be considered. There is, however, no doubt that sorghum domestication has induced a significant reduction of its molecular diversity. Considering a sample that is representative of the extensive diversity of sorghum together with a whole genome sequencing approach, [49] showed that nucleotide diversity estimated through $\square_{\pi}$ and $\square_{W}$ were respectively 35% and 28% lower in sorghum landraces compared to the wild genotypes. These reductions reach respectively 39% when considering the whole genome and 34% when considering the genic regions only. The present paper aims at studying whether or not this documented loss of allelic diversity is accompanied by a loss of diversity in gene isoform relative expression.

The growing evidence of widespread intraspecific variability of AS, along with its potential role in adaptation makes it susceptible to demographic and selective events. As plant domestication is a well-studied evolutionary process, during which demographic and selective effects are combined, we ask if, and how, domestication may have impacted AS. We ask also whether an extreme difference of AS patterns, between wild and cultivated accessions for a given gene, could be the signature of a selective effect on this gene AS pattern itself. Taking advantage of an mRNA dataset produced to document the domestication of several agronomical species [50] we chose to focus on sorghum for the quality of its genome assembly and annotation. To supplement [50] and [51] we used an additional sorghum accession (WS7) to be able to balance the number of accessions so that we had ten for each compartment. RNAseq data from these ten cultivated and wild sorghum accessions were mapped on the sorghum reference genome. We focused on genes for which exactly two isoforms were identified and we studied the variability of the expression ratio between those isoforms across compartments.

Material and methods

Sorghum genome and annotation

We used the sorghum genome assembly Sbi1.4 and the corresponding transcript annotations provided on the plantGDB database (http://www.plantgdb.org/XGDB/phplib/download.php?GDB=Sb). The gene ontology annotations of those annotated sorghum genes have been downloaded thanks to the biomart facilities of the plant ensEMBL database.

Biological material

Ten accessions of cultivated sorghum have been used to produce the sequence information, *Sorghum bicolor* subsp. *bicolor* (denoted CS1 to CS10), and ten wild relatives (denoted WS1 to WS10), chosen in order to best represent the genetic diversity of each compartment (Table 1). Note that below we used indifferently the terms ‘population’ and ‘compartment’.

We were mainly interested in comparing features observed within the compartment of 10 cultivated sorghum accessions, denoted as popCS$_{10}$ below, with those observed in the sample of 10 wild sorghum accessions, denoted as popWS$_{10}$.

Table 1. Accession names and origins of sequenced sorghum accessions.

Sorghum bicolor bicolor (Cultivated sorghum: CS)			*Sorghum bicolor verticilliflorum* (Wild type sorghum: WS)		
Study code	**Accession**	**Country**	**Study code**	**Accession**	**Country**
CS1	SSM1049	Senegal	**WS1**	IS14564	Sudan
CS2	IS29876	Swaziland	**WS2**	IS18821	Egypt
CS3	IS30436	China	**WS3**	IS18909	Chad
CS4	SSM1123	Niger	**WS4**	IS18824	Ivory Coast
CS5	IS6193	India	**WS5**	IS18833	Malawi
CS6	SSM973	Senegal	**WS6**	IS14312	South Africa
CS7	IS14317	Swaziland	**WS7**	IS14357	Malawi
CS8	IS29407	Lesotho	**WS8***	IS14719*	Ethiopia
CS9	SSM1057	Senegal	**WS9**	IS18804	USA
CS10	IS26554	Benin	**WS10**	IS18812	Egypt

* This accession was mis-assigned to the wild compartment (see next paragraph in M&M section).

https://doi.org/10.1371/journal.pone.0183454.t001

Preliminary genomic analysis raised doubts concerning the assignation of the accession WS8 as a wild type. Indeed, SSR verifications and phenotypic observations of the seed lot received from the genebank revealed a misidentification. Additionally, a surprisingly low percentage of reads from accessions WS1, WS2 and WS5 could be properly mapped on the reference sorghum genome (details in result section). Thus, we removed those 4 accessions from our initial wild type sample $popWS_{10}$ (thereby generating a sample we noted $popWS_6$) and, to check for potential bias induced by sample sizes, we randomly subsampled 6 accessions in the cultivated sample. Four such subsamples were obtained (called $popCS_{6\_1}$, $popCS_{6\_2}$, $popCS_{6\_3}$, $popCS_{6_4}$below).

We use $popCS_x$ (respectively $popWS_x$) to designate one of the above mentioned samples of cultivated sorghum (respectively wild sorghum) in assertions that hold for all of cultivated (respectively wild type) samples. Finally, we use $popS_x$ to designate any of those sorghum samples.

RNA extraction and sequencing

The RNAseq data used were obtained from a larger project dedicated to the comparison of cultivated plants with their wild relatives (http://www.arcad-project.org/projects/comparative-population-genomics). Tissue samples were collected from different organs, including leaves, grains, and inflorescence. Details for RNA extraction, Illumina libraries production and sequencing conditions are available in the Materials and Methods section of [50]. The cDNA libraries that contain a mixture of 65% RNA from the inflorescence, 15% from leaves and 20% from maturing seeds, for each accession, were sequenced using the Illumina mRNA-Seq, paired-end protocol on a HiSeq2000 sequencer (one run for each compartment). The paired-end reads, in the illumina FASTQ format, were cleaned using cutAdapt [52] to trim read ends of poor quality (q score below 20) and to keep only those with an average quality above 30 and a minimum length of 25 base pairs. Those data are freely available on the NCBI RSA database (Sequence Read Archive) (cultivated: SAMN05277472 to SAMN05277481; wild: SAMN06052464 to SAMN06052472 and SAMN07313361).

Estimation of alternative transcript expression levels

Transcript expression levels have been estimated thanks to the Tuxedo pipeline [12]. This pipeline proceeds as follows. Firstly, for each accession, RNAseq reads are mapped on the reference genome using Tophat v2.0.13 [53] with bowtie2 v2.2.5 [54]. Secondly, the resulting mappings are used to enrich the initial gene and transcript predictions used, thanks to cuffmerge and cufflink, two programs of the cufflink suite v2.2.1 [55]. Finally, reads mappings and enriched annotations are combined to estimate, for each gene and accession, the expression level of every alternative transcript using cuffdiff, another program from the cufflink suite. The expression level is measured by cufflink as an 'FPKM' (Fragments Per Kilobase Of Exon Per Million Fragments Mapped), to account for heterogeneity of i) total number of reads per individual and ii) mRNA length.

When the average depth coverage of a gene was smaller than 5 for an accession, we considered that the corresponding expression level could not be reliably estimated and we replaced the cuffdiff estimation by a missing data (NA) for the corresponding gene in the considered accession.

Estimation of alternative transcript expression ratios

We compare two panels of genotypes, popWS$_x$ and popCS$_x$, based on a subset of genes selected according to the following characteristics: i) genes expressing exactly two alternative transcripts (6,226 genes taken from the 33,795) ii) genes having an average depth coverage of at least 5 reads for every accession of popWS$_x$ and popCS$_x$ (*i.e.* no missing data) and iii) transcripts of genes both being expressed in at least one accession of popWS$_x$ and at least one accession of popCS$_x$. These filters, being quite stringent, still allow us to rely on more than a thousand genes for comparing any pair of wild/cultivated samples (cf. Results section). For such genes with exactly two isoforms, the alternative transcript expression levels can be summarized by a single expression ratio, denoted as e$_T$-ratio below. The e$_T$-ratio is simply the expression of one transcript divided by the overall expression of the gene. For a given gene, if we denote by x its e$_T$-ratio then using the alternative transcript at the numerator would have led to an e$_T$-ratio of 1-x. As long as the same isoform is used to calculate the e$_T$-ratio for all accessions (within the cultivated and wild samples), using one isoform or the other at the numerator of a gene e$_T$-ratio does not matter when comparing their diversity in cultivated versus wild type samples. To homogenize the presentation of the results among genes we therefore systematically used, for the e$_T$-ratio numerator of a gene, the isoform leading to the highest average e$_T$-ratio along popWS$_x$ U popCS$_x$, so that most of our e$_T$-ratios range between 0.5 and 1 instead of being evenly spread between 0 and 1.

Estimation of transcript expression diversity within population

For a given gene G and sample popS$_x$, the diversity of the transcript expression is simply the diversity of its e$_T$-ratios among the considered sample. If all e$_T$-ratios of the given sample are close to 1, the 'first' transcript of G (*i.e.* the isoform which, on average, is the most expressed and hence used as the numerator of the e$_T$-ratio) is much more expressed than its alternative transcript in all accessions of this population. Note that e$_T$-ratios can be roughly constant among accessions of popS$_x$ no matter the value of this constant. The diversity of the expression balance between the two isoforms of gene G among popS$_x$ can be measured by the spread of its e$_T$-ratios, which can be quantified using either their variance (denoted as σ_r) or their inter-quartile range (denoted as iq$_r$). Both measures capture the variability of the e$_T$-ratios but the variance is much more sensitive to outlier e$_T$-ratio values than the inter-quartile range. Similarly, we will summarize the e$_T$-ratios of a gene G among the popS$_x$ using either the average (denoted as avg$_r$) or the median (med$_r$) of the e$_T$-ratios of G in popS$_x$.

Results

Dataset characteristics

For each accession, the proportion of clean paired-end reads that successfully mapped on the sorghum V1.4 genome is provided in Fig 1. Less than 50% of the clean reads of individuals WS1 (37.8%), WS2 (46.6%) and WS5 (46.7%) have been successively mapped on the sorghum genome. This low percentage strongly contrasts with other accessions for which at least 81.3% (for individual WS3) of the read pairs have been successively mapped. Similar results were obtained with other mapping tools, showing that this is not just an artifact of the chosen mapping method. We did not find any satisfactory explanation to this low percentage of read mapping and preferred to discard those three accessions for the current analysis together with individual WS8 for which we have some suspicions of misidentification.

The impact of the gene filtering applied to our dataset, in order to base our population comparisons solely on genes with a relatively high sequencing coverage and no missing data in the compared populations, is detailed in Table 2. Note that despite quite a drastic filtering procedure, all pairwise population comparisons are conducted on more than one thousand genes.

Among the 1397 genes harboring exactly two isoforms (comparison popCS6_2 vs popWS6 Table 2, the highest number of genes among all the comparisons), 826 genes were already identified with two isoforms of transcripts in the publically available annotation, 556 genes

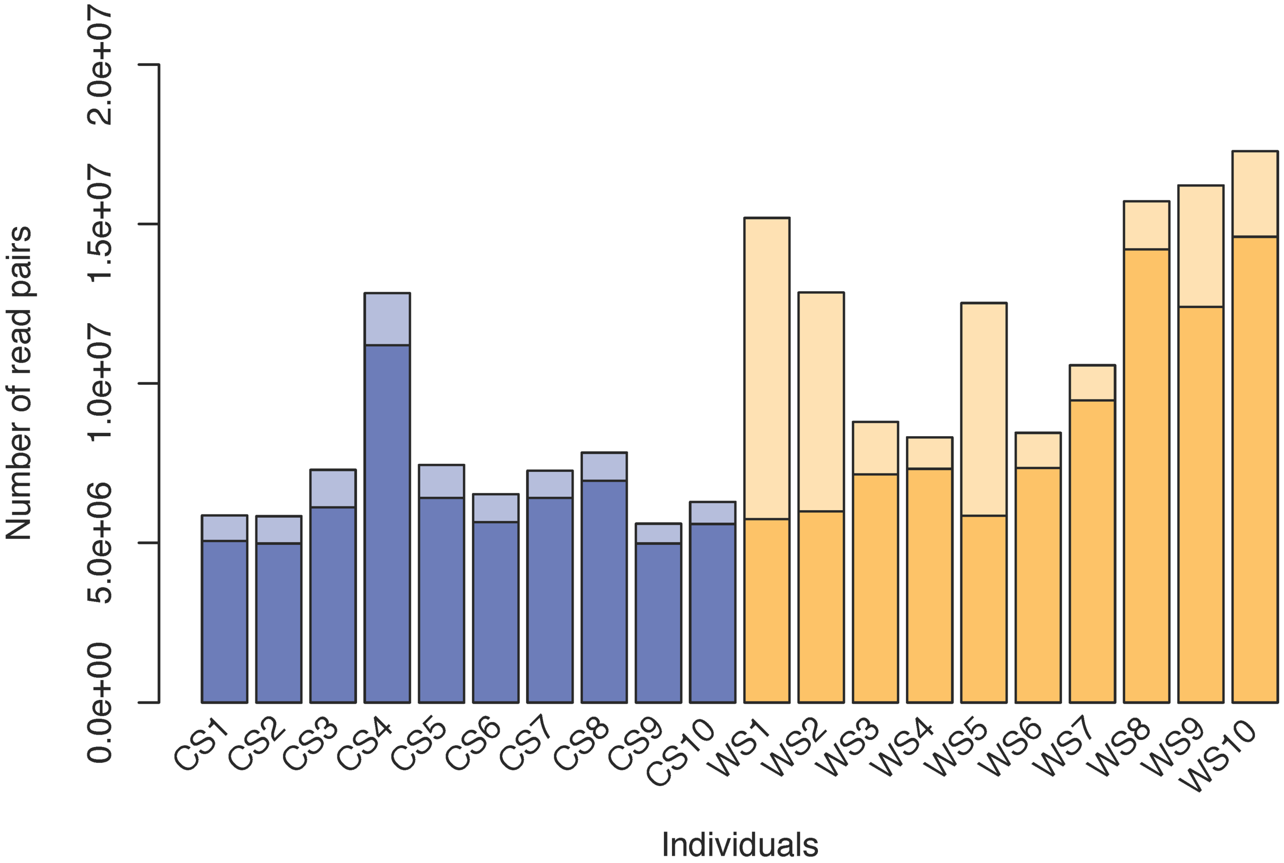


Fig 1. Number of clean pairs of reads mapped on the sorghum genome. The number of clean read pairs of each individual is indicated by a blue bar for cultivated sorghum accessions or an orange bar for wild sorghum accessions. For any given accession, the darker or lighter part of each bar corresponds to mapped or not mapped read pairs on the sorghum genome.

https://doi.org/10.1371/journal.pone.0183454.g001

Table 2. Number of genes considered for the analysis after filtering on quality and coverage.

Compared Population	Number of genes with 2 isoforms	and a gene coverage above 5 for each individual	and both isoforms expressed in both populations
popCS$_{10}$ vs popWS$_{10}$	6,226	1,385	1,134
popCS$_{10}$ vs popWS$_{6}$	6,226	1,635	1,358
popCS$_{6\_1}$ vs popWS$_{6}$	6,226	1,653	1,350
popCS$_{6\_2}$ vs popWS$_{6}$	6,226	1,698	1,397
popCS$_{6\_3}$ vs popWS$_{6}$	6,226	1,668	1,356
popCS$_{6\_4}$ vs popWS$_{6}$	6,226	1,682	1,383

https://doi.org/10.1371/journal.pone.0183454.t002

were annotated with only one isoform of transcript, and 15 genes correspond to loci where no genes were identified. The information related to these genes is available in S1 Table (gene id, protein sequence when predictable, its length) and includes the nucleotide identifier number for mRNA sequences available in S1 File.

Distribution of e_T-ratio within cultivated and wild type sorghum

For each pairwise population comparison, we used either e_T-ratio mean values and variances within each population (Table 3), or e_T-ratio medians and interquartiles (Table 4). For all popCS$_x$ vs popWS$_x$ comparisons, diversity of e_T-ratios is significantly higher in cultivated populations than in domesticated ones. Indeed, most genes have an e_T-ratio variance higher in the wild population than in the cultivated one. For instance, e_T-ratio variance is higher in popWS$_{10}$ than in popCS$_{10}$ for 773 genes out of 1134 (~68%). The percentage of genes having an e_T-ratio which is more variable in the wild population than in the cultivated population varies depending on the compared populations but is always significantly higher than 50% according to paired student t-test (highest p-value $1.63e^{-110}$) and Wilcoxon test (highest p-value $5.09e^{-10}$). The same observation holds true for comparisons based on e_T-ratio medians and inter-quartile ranges. In all population comparisons but one, the inter-quartile range is very significantly lower in the cultivated population (p-value$<1.50e^{-8}$ for student test and $<1.79e^{-9}$ for Wilcoxon test). The sole minor exception is for the comparison of popCS$_{10}$ and

Table 3. Comparison of the e_T-ratios variance between cultivated and wild sorghum samples.

	popCS$_{10}$ popWS$_{10}$	popCS$_{10}$ popWS$_{6}$	popCS$_{6\_1}$ popWS$_{6}$	popCS$_{6\_2}$ popWS$_{6}$	popCS$_{6\_3}$ popWS$_{6}$	popCS$_{6\_4}$ popWS$_{6}$
# σ_r (CS) > σ_r (WS)	347	599	545	592	548	532
# σ_r (CS) = σ_r (WS)	14	1	14	19	24	15
# σ_r (CS) < σ_r (WS)	773	758	791	786	784	836
slope of linear regression of (σ_r (CS), σ_r (WS))	0.5228	0.5730	0.6326	0.5393	0.5871	0.5609
Paired t-student						
mean(σ_r (CS) - σ_r (WS))	0.0058	0.0021	0.0020	0.0021	0.0020	0.0028
p-value of t-student	$3.63e^{-119}$	$4.46e^{-142}$	$2.40e^{-120}$	$1.63e^{-110}$	$1.23e^{-111}$	$7.80e^{-112}$
p-value Wilcoxon test	$1.48e^{-52}$	$5.09e^{-10}$	$4.14e^{-14}$	$6.38e^{-13}$	$2.14e^{-14}$	$5.99e^{-22}$

Each column corresponds to the comparison between a sample of cultivated genotypes and a sample of wild genotypes. In the first (resp. second and third) line are reported the number of genes with an e_T-ratio variance (σ_r) in the cultivated panel higher than (resp. equal to, lower than) in the wild sample. The fourth line indicates the slope of the linear interpolation of the points having σ_r (CS) as abscises and σ_r (WS) as ordinate. The mean value of the differences between σ_r (CS) and σ_r (WS) is provided in the next line, and the last two lines provide respectively the p-value of the paired t-test and the p-value of the Wilcoxon test to statistically asses the significance of the difference between σ_r (CS) and σ_r (WS) distributions.

https://doi.org/10.1371/journal.pone.0183454.t003

Table 4. Comparison of the e_T-ratio inter-quartile range between cultivated and wild sorghum samples. (see detailed legend in Table 3).

	popCS$_{10}$ popWS$_{10}$	popCS$_{10}$ popWS$_6$	popCS$_{6\_1}$ popWS$_6$	popCS$_{6\_2}$ popWS$_6$	popCS$_{6\_3}$ popWS$_6$	popCS $_{6_4}$popWS$_6$
# iq$_r$ (CS) > iq$_r$ (WS)	379	635	545	569	540	550
# iq$_r$ (CS) = iq$_r$ (WS)	80	60	59	67	63	57
# iq$_r$ (CS) < iq$_r$ (WS)	675	663	746	761	753	776
slope of linear regression of (iq$_r$ (CS), iq$_r$ (WS))	0.4714	0.6013	0.5402	0.4931	0.5058	0.4829
Paired t-student						
mean(iq$_r$ (CS)—iq$_r$ (WS))	0.0284	0.0060	0.0140	0.0132	0.0158	0.0175
p-value of t-student	7.44e^{-25}	0.0039	1.27e^{-9}	1.50e^{-8}	4.09e^{-12}	1.67e^{-13}
p-value Wilcoxon test	6.24e^{-29}	0.0452	9.25e^{-11}	1.79e^{-9}	2.05e^{-12}	4.31e^{-15}

Each column corresponds to the comparison between a sample of cultivated genotypes and a sample of wild genotypes. In the first (resp. second and third) line are reported the number of genes with an e_T-ratio inter-quartile range (iq$_r$) in the cultivated panel higher than (resp. equal to, lower than) in the wild sample. The fourth line indicates the slope of the linear interpolation of the points having iq$_r$ (CS) as abscises and iq$_r$ (WS) as ordinate. The mean value of the differences between iq$_r$ (CS) and iq$_r$ (WS) is provided in the next line, and the last two lines provide respectively the p-value of the paired t-test and the p-value of the Wilcoxon test to statistically asses the significance of the difference between iq$_r$ (CS) and iq$_r$ (WS) distributions.

https://doi.org/10.1371/journal.pone.0183454.t004

popWS$_6$, two populations of different sizes, that do have significantly different e_T-ratio inter-quartile ranges but with not so low p-values (p-value 0.0039 for the paired student t-test and 0.0452 for the Wilcoxon test). The simple e_T-ratio dot plot displayed in Fig 2 gives visual prominence to this general trend of higher variance (or interquartile range) of e_T-ratio in wild populations than in cultivated ones.

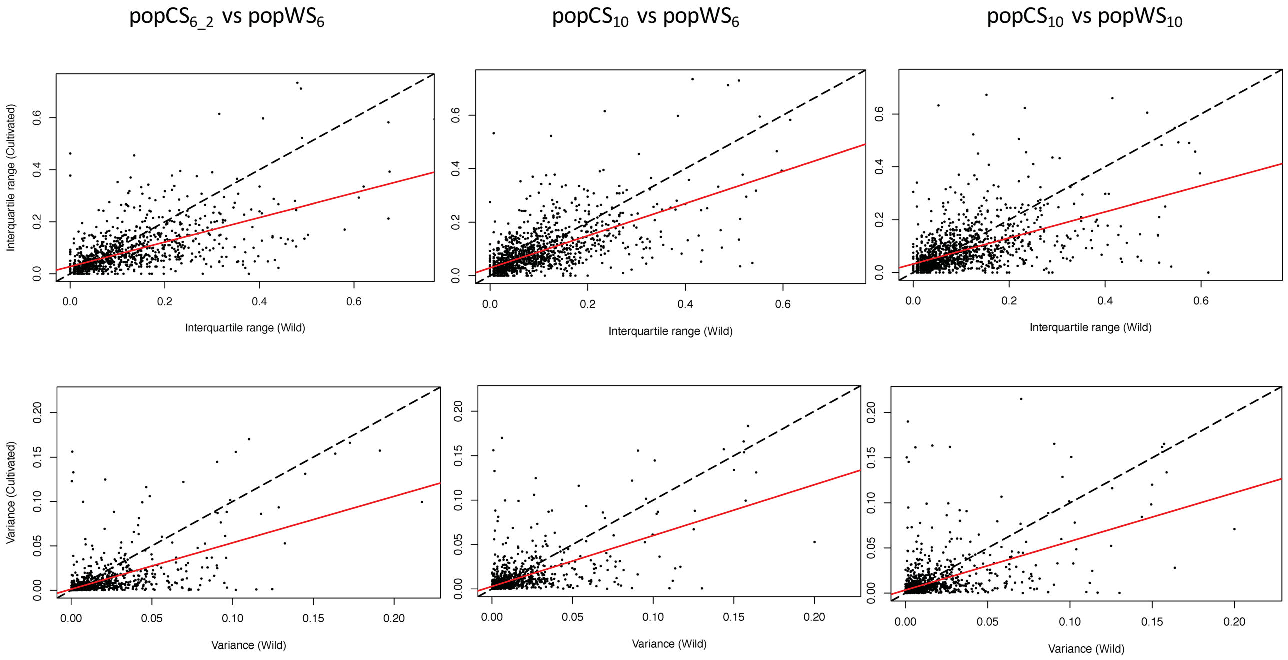


Fig 2. Dot plot comparison of the e_T-ratio spread among cultivated and wild type populations. In each plot a dot represents a gene whose position corresponds to its e_T-ratio spread measure by interquartile range (resp. variance) in the three top (resp. bottom) plots, observed in a sample of cultivated sorghum (abscise) and in a sample of wild sorghum accessions (ordinate). The red lines represent the linear interpolation of those points (the line slopes are provided in Tables 3 and 4) and the dashed lines depict the y = x line to ease picture interpretation.

https://doi.org/10.1371/journal.pone.0183454.g002

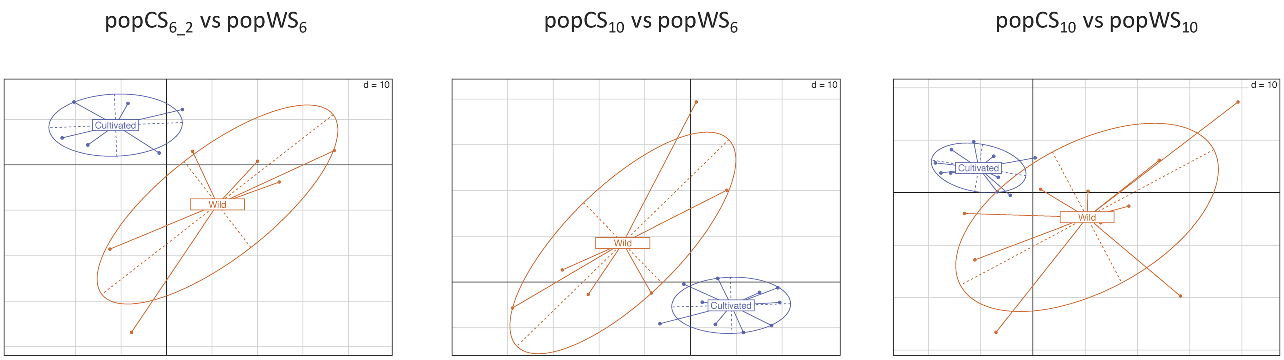


Fig 3. Cultivated and wild type sorghum sample 2D projection using a PCA of their e_T-ratios. Each sorghum accession, associated with e_T-ratios, can be seen as a point in a high dimensional space. This figure displays the projection of these points on the two first PCA axes using orange /blue dots to represent wild /cultivated individuals. The two first axes explain more than 30% of the original variability in all three cases.

https://doi.org/10.1371/journal.pone.0183454.g003

Organization of sorghum accessions based on their e_T-ratios

The e_T-ratios are not only less variable in the domesticated compartments, they also seem to be sufficient to correctly differentiate cultivated accessions from wild type accessions. Considering the e_T-ratio of each gene as a coordinate, each accession can be positioned in a highly multidimensional space. The usual Principal Component Analysis (PCA) can then be used to project these accessions/points in a lower dimensional space while preserving most of the original variance. The projection obtained on the two first axis of the PCA analysis are provided in Fig 3 where cultivated accessions group together in a much more compact group than the wild individuals. Note also that, in the three PCA projections displayed in Fig 3, the two axes used for the projection explain more than 30% of the original variance.

Genes with contrasted e_T-ratios distribution in wild and cultivated sorghum

Genes with contrasted e_T-ratio variability between the cultivated and wild compartments are potentially related to the domestication syndrome. To identify such genes, we were looking for genes having an interquartile range which differs between both populations by at least 0.2, *i.e.* genes such as $|iq_r\,(CS)—iq_r\,(WS)| > 0.2$. We found about twice as many genes with a higher interquartile range in wild compartments compared to the cultivated ones, than the opposite way around (Fig 4). All identified genes are interesting as such contrasts of e_T-ratio, whatever their orientation, may reveal genes that have been affected by domestication.

A total of 59 genes were identified when comparing $popCS_{6\_2}$ and $popWS_6$ (Fig 4), among which nineteen are consistently recovered by the three population comparisons we focus on. Twelve out of these nineteen genes are annotated by specific GO terms. To find out if some GO terms are over represented in this set of 12 genes with respect to the set of 921 annotated genes common to the three population comparisons, we relied on the AgriGO webserver (http://bioinfo.cau.edu.cn/agriGO/analysis.php). This enrichment analysis was done using Singular Enrichment Analysis (SEA) with hypergeometric test, p-value threshold at 0.05 and Bonferroni correction for multiple testing. A single annotation is found to be over-represented by this test (p-value 0.00042), the GO term 'regulation of biological quality' (GO:0065008). This GO term has a frequency of 0.42 (5/12) in our subset versus a frequency of 0.04 (40/921) in the subset of annotated genes common to the 3 population comparisons. The five genes

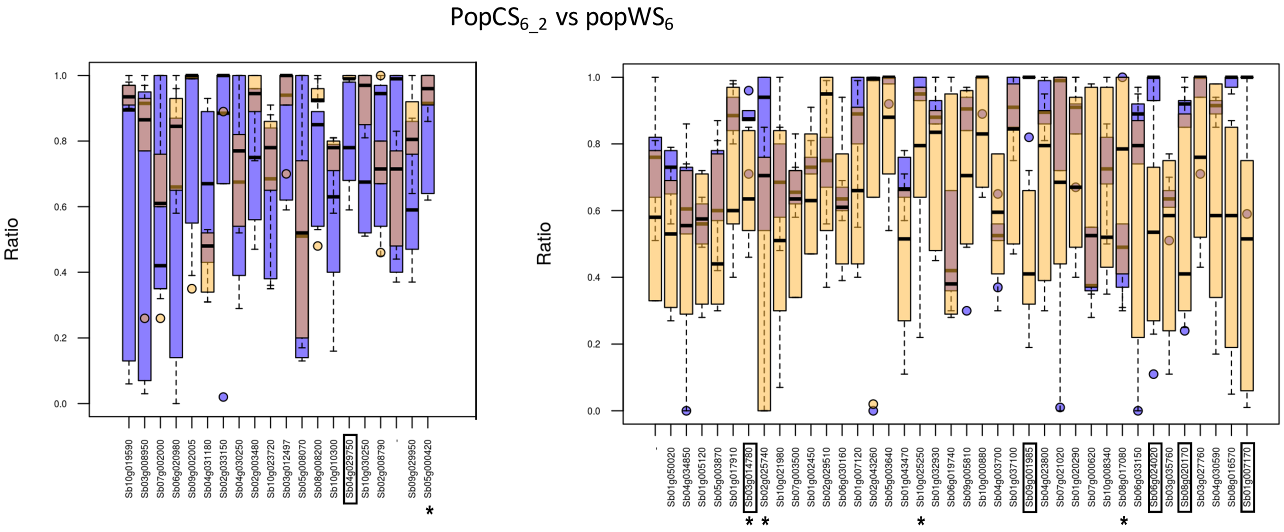


Fig 4. Genes with a contrasted e_T-ratio interquartile in cultivated (popCS$_{6\_2}$) and wild (popWS$_6$) samples. Box plot representations of the e_T-ratio in the cultivated (blue) and wild (orange) samples, for genes with an e_T-ratio more variable in cultivated (left) or wild (right) samples. The genes marked by a star are annotated by the GO term 'regulation of biological quality'. The framed genes are common to Fig 5.

https://doi.org/10.1371/journal.pone.0183454.g004

annotated by this GO term are Sb02g025740, Sb03g014780, Sb05g000420, Sb08g017080, and Sb10g025250 (marked by a star in Fig 4).

Finally, we were looking for genes having an e_T-ratio (isoform expression balance) that strongly differs in wild and cultivated population. More precisely, we were searching for genes with a difference of e_T-ratio median value in cultivated and wild compartments greater than 0.2. For this filter, we added the constraint that the median difference should also be superior to the average intrapopulation e_T-ratio spread, leading to the following filter formulation: $|\mathrm{med}_r(CS)\text{- }\mathrm{med}_r(WS)| > \max(0.2, (\mathrm{iq}_r\ (CS) + \mathrm{iq}_r\ (WS))/2)$. This provides us with genes that have a difference in e_T-ratio between the two compartments (cultivated / wild) that exceed differences observed within compartments (Fig 5). A total of 47 genes were identified when comparing popCS$_{6\_2}$ and popWS$_6$, among which fifteen were common to the three population comparisons we are focusing on, but this time, we found no over represented GO-term among these genes.

Six genes are common to both comparisons and are contrasted between the cultivated and wild compartments for both e_T-ratio interquartile and median: Sb01g007170, Sb08g020170, Sb03g014780, Sb06g024020, Sb09g001985 and Sb04g029750. These genes are framed in Figs 4 and 5.

We then tried to estimate the potential impact of the AS events for those genes, by comparing the 'alternative' protein to the canonic one predicted according to the annotation of each gene in the sorghum genome version Sbi1.4 (cf. M&M section). The AS events were classified into 4 categories. In the first category, the start codon of the canonic form is not present anymore (no translation prediction is made). In the second category, a stop codon appeared very early (in the first 20% of canonical protein), often due to an early frame shift. In these two cases we can speculate that the AS event may be deleterious (although it can have a role in mRNA degradation). In the third category, the alternative protein is slightly different from the canonical one (*i.e.* either with indels affecting less than ten percent of the protein, or identical on more than 50% of the protein but with an equivalent length, or with an identical sequence on a minimum length of 500 amino acids). In the last category the two proteins are 100%

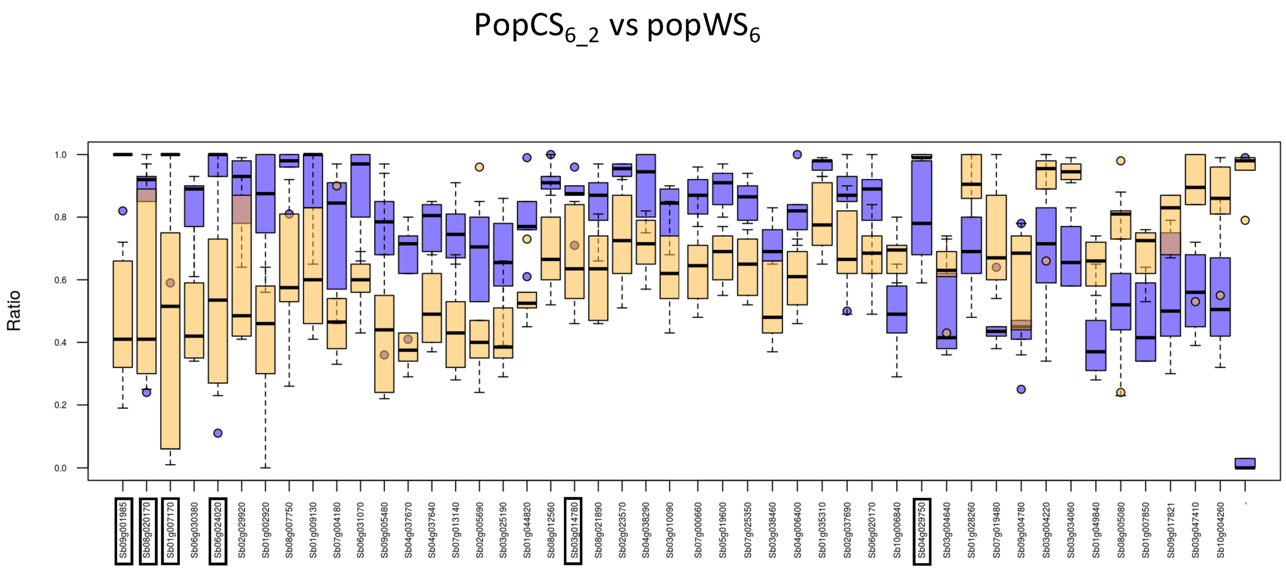


Fig 5. Genes with a contrasted e_T-ratio median in cultivated (popCS$_{6\_2}$) and wild (popWS$_6$) samples. Box plot representations of the e_T-ratio in the cultivated (blue) and wild (orange) sample. The framed genes are common to Fig 4.

https://doi.org/10.1371/journal.pone.0183454.g005

identical (*i.e.* the AS event concerned only UTR). We can speculate that the alternative protein isoform is functional in these two last categories and that the equilibrium between both mRNA isoforms may have a biological significance (either regulation of amounts of protein, or different roles of the protein themselves). Table 5 provides the distribution, in the 4 above mentioned categories, of the genes having a contrasted e_T-ratio interquartile in cultivated and wild compartments (genes detailed in Fig 4).

Finally, in order to go further with the interpretation of our results, we were trying to determine the function of the genes identified as having a contrasted e_T-ratio distribution, and, in particular, to look for genes potentially involved in traits related to the domestication syndrome.

As mentioned above, Sb03g014780 belongs to the GO term 'regulation of biological quality' which was over-represented in the gene-set presenting the highest e_T-ratio interquartile contrast (Fig 4). This gene is also identified as having a high e_T-ratio median difference between the wild and the cultivated compartment (Fig 5, framed). The protein predicted for this gene presents 96% of identity with the protein accession Q7G8Y3.2 encoded by the rice gene Os01g0367900. This protein corresponds to a probable chromatin-remodeling complex ATPase chain also known as ISW2 (Imitation Switch Protein 2) which is involved in coordinating transcriptional repression in saccharomyces cerevisiae [56]. The alternative isoform is lacking 28 amino acids, located in a region where three nucleotide binding sites are detected. The deletion is located precisely between the two last nucleotide binding sites, resulting in the merging of those sites. Consequently, in the alternative protein isoform only two nucleotide

Table 5. Distribution of genes with contrasted e_T-ratio interquartile in cultivated and wild compartments (Fig 4) according to the potential functional impact of the alternative isoform.

Genes with larger e_T-ratio interquartile in	Identical protein	Potentially functional protein	Total 'functional'	'Non functional' protein	No protein identified	Total 'deleterious'
Cultivated (Fig 4 left)	8	10	**18 (95%)**	0	1	**1 (5%)**
Wild (Fig 4 right)	10	10	**20 (53%)**	2	16	**18 (47%)**

https://doi.org/10.1371/journal.pone.0183454.t005

binding sites are detected. Experimental data would be needed to investigate if its efficiency is affected as it can be predicted from the in silico analysis.

Two other genes annotated by the above-mentioned GO term 'regulation of biological quality' and having contrasted e_T-ratio interquartile, are homologous to genes identified in selection scan studies or genome wide association studies (GWAS) in other species, underlying their putative impact on plant phenotype.

The first one is Sb02g025740. The protein predicted from this gene presents 69% identity with the protein accession Q8LCQ4.1 which is encoded by the LHCA6_ARATH locus from *Arabidopsis thaliana* (At1g19150). This protein corresponds to a Photosystem I light harvesting chlorophyll a/b also known as Light Harvesting Complex. These proteins, through their interactions with the core complexes of both photosystems, are involved in the enhancement and regulation of light-harvesting, the transfer of light energy to the photosynthetic reaction centers and also provide protection against photo-oxidative stress [57]. Photosystem Light harvesting chlorophyll a/b proteins have been identified through genome wide association studies as being involved in the photochemical reflectance index in Soybean [58] and to several agronomical traits (height, spike length, number of grains per spike, thousand grain weight, flag leaf area and leaf color) in barley [59]. The alternative splicing event identified for this gene is showing an insertion of only one amino acid in position 16, the rest of the protein is 100% identical to the canonic form.

The second gene is Sb10g025250. Its derived protein presents 73% identity with the protein accession Q949Y3 encoded by At5g34850. This protein corresponds to a bifunctional purple acid phosphatase. Purple acid phosphatases are known to be involved in phosphate acquisition and play a role in phosphate deficiency adaptations [60, 61]. In a recent study on soybean, the gene *GmACP1* was identified as playing a significant role in soybean tolerance to low phosphorus [62]. In addition, in *Helianthus annuus*, evidence of selective sweeps combined with higher than expected Fst values were also identified for a purple acid phosphatase [63].

The last gene identified with a high e_T-ratio interquartile difference (Fig 4) and for which a function can be predicted is Sb04g030590. This gene codes for a protein showing 93% identity with a soluble inorganic pyrophosphatase (Q0DYB1) encoded by the rice gene Os02g0704900. This protein catalyzes the irreversible hydrolysis of pyrophosphate [64]. In apple, one locus showing signature of selection between wild and domesticated apples was located in a gene coding for an inorganic pyrophosphatase, and this function is described as associated with sugar metabolism and acidity [65]. Indeed, Fruit quality traits have played critical roles in domestication of the apple [65].

Finally, among genes for which a high difference of e_T-ratio median value is observed between the wild and cultivated sample (Fig 5), the gene Sb1g007850 is potentially involved in 'the flowering pathway', another trait of agronomic interest which is often mentioned as a target of the domestication process. Indeed this gene presents more than 90% of amino acid identity with the photoreceptor phytochromes C, from several grass species including rice, maize and *Brachypodium dystachion*. In the temperate model grass *Brachypodium dystachion*, phytochromes C has been shown to be an essential light receptor involved in photoperiodic flowering [66]. In pearl millet, natural variations at the phytochrome C locus are linked to flowering time and morphological variations [67]. The alternative isoform detected with our RNAseq data does not comprise the start codon of the canonical form. Only one copy of phytochrome C is identified in sorghum and it is tempting to speculate that the alternative isoform may be deleterious. The e_T-ratio between both forms is clearly different between the wild and the cultivated compartment (Fig 5). However, drawing conclusions about a potential selective effect at this locus, linked to domestication would require additional investigations.

Discussion

Domestication has been shown to impact phenotypic traits, genetic diversity, and gene expression and to be associated with selective effects on a wide number of loci. One study comparing AS profiles in domesticated maize and its wild relative teosinte, has recently been published [39]. To our knowledge, this is the sole publication comparing AS between wild and cultivated plants, and nothing at all has been published so far regarding the impact of domestication on the relative expression of gene isoforms or, more generally, on the diversity of AS expression levels. Here we relied on available RNAseq data to document AS expression variability between wild and cultivated sorghum.

Strict filters are needed to focus on 'non-erratic' AS events

The biological meaning of the complex splicing landscape is still not totally understood. Within the population of mRNA molecules, some variants are issued from random splicing errors and can be assimilated to background noise. Those erratic AS events are not supposed to be present in high frequency. They can therefore be eliminated, or at least strongly minimized, by increasing the sequencing coverage threshold used to assert the presence of isoforms. The remaining AS events may be qualified as 'non-erratic' AS events and may have a positive, neutral or negative impact on the organism. They are, somehow, controlled and induced by genetic and/or environmental factors and should be, at least partially, heritable. As such, they are expected to be consistently found in a given genotype, and potentially in other genotypes of the same species, provided that sequencing and environmental conditions are similar.

Our analyses rely on a subset of genes expressing exactly two isoforms. We applied several filters to remove as many as possible of the erratic AS isoforms *i.e.* a minimum coverage (depth of sequencing) over the whole transcript and isoform presence in at least two individuals (one cultivated and one wild). According to the high level of expression of the genes we selected in our dataset, and the observed consistency among wild and cultivated compartments, we are quite confident that the AS events we were focusing on are not background noise of splicing machinery.

The domestication bottleneck is most likely the cause of the global reduction of e_T-ratio variability in domesticated sorghum

A strong and significant loss of variability of e_T-ratio (*i.e.* balance of the two isoforms resulting from AS) is observed between wild and cultivated compartments (Tables 3 and 4, Fig 2). This result is observed irrespective of the accession samplings considered for each compartment and thus extremely reliable. If domestication has been shown to substantially reduce nucleotide diversity in a vast range of species, including sorghum [45, 49], we show here, for the first time, that domestication also impacts the regulation of the alternative splicing process itself.

AS regulation appears extremely complex and sensitive to environmental stimuli [24]. In this study, the mRNA extraction conditions were—as much as possible—homogenous for all genotypes, we assume that the variability observed is mainly reflecting the genotypic variability. A parallel can be drawn between our results and the results obtained in beans for which a very clear decrease of gene expression variability (18%) was also detected in domesticated beans as compared to their wild counterparts [36]. This loss of expression variability was interpreted as a direct consequence of the strong loss of genetic diversity observed during common bean domestication (almost 50% in coding sequences) affecting DNA regions involved in transcription regulation. Here we assume that the significant loss in AS variability we observe in sorghum is also due to nucleotide variability loss during domestication, in particular in regions where both *trans* and *cis* elements controlling AS are located. Two additional elements reinforce this hypothesis. First, in maize, the diversity of *cis* regulatory elements has been shown to be reduced by domestication and the *cis* element themselves have been suggested to be targets of selection during domestication [68]. Second, in humans, several studies show that AS is, at least partially, controlled by nucleotide diversity present in genomic regions which are more or less close to the target gene [69, 70, 71]. A reduction of nucleotide variability in these regions, whatever their exact distance to the targeted genes, is expected to impact AS variability. Genome wide association mapping on AS variability using either wild or domesticated plants could help to further document these interactions.

The global loss of e_T-ratio variability, observed between the cultivated and the wild sorghum compartments, is most likely due to the loss of nucleotide diversity induced by the strong demographic bottleneck caused by domestication. Under this neutral, genetic drift related, assumption, the balance between isoforms is expected to have a minute effect for most genes. However, the cumulative effect over the genome might be an important component of the genetic load incurred by domestication. Results from Table 5 tend to confirm this hypothesis.

Indeed, AS events are approximately equally distributed between 'functional' and 'deleterious' in genes for which the e_T-ratio interquartile is higher in the wild compartment, in agreement with the neutral hypothesis of this expression diversity reduction. Note that, though the global trend of AS annotation provided in Table 5 may be informative, each individual AS annotation should not be taken for granted. The assignation of a specific AS event as leading to either a functional or non-functional protein needs to be empirically confirmed. Indeed, although an early stop codon is a gage of loss of protein functionality, the effect of the other mutations is not as easily predictable.

The e_T-ratios may provide valuable insight for a better understanding of the domestication syndrome

The extent of the nucleotide diversity loss due to domestication is used to characterize the strength of the demographic bottleneck occurring during the domestication process itself [30]. In sorghum, the strength of the bottleneck has been documented to be around 25% (θ_π) and 38% (θ_W) at the whole genome level [49]. When only genic regions are considered the strength of the bottleneck is estimated around 39% (θ_π) and 34% (θ_W) [49]. The slopes of the linear regressions between wild and cultivated e_T-ratio are between 0.52 and 0.63 for e_T-ratio variance and between 0.47 and 0.60 for e_T-ratio inter-quartile range (Tables 3 and 4). These values could be seen as another insight of the intensity of the bottleneck but are much higher than those derived from nucleotide polymorphism studies. Although it is hazardous to compare these values (different methods and slightly different datasets) we can conclude that the impact of domestication on AS is strong. It is also possible that the nucleotide diversity reduction impacted some loci with pleiotropic effects on AS regulation. The impacts of domestication on AS would deserve to be explored in other species in order to determine whether such a large impact is specific to sorghum or if it is a general trend among domesticated species.

After having discussed the fact that the global decrease of e_T-ratio variability in the domesticated compartment can be interpreted as a consequence of demographic bottlenecks, we now ask whether this result is entirely neutral (affected by demographic events only), or if it could also result from selective effects. In other words, may a given isoform, or ratio between two isoforms, have increased in frequency in the cultivated compartment because it procures an advantage in the domesticated context, as found for key genes controlling the domestication syndrome [30, 31]. At the genome scale, domestication tends to reduce diversity, however, a gain of diversity can be locally associated with the post domestication diversification especially for loci responsible for interesting traits. This possibility is supported by the results provided in Table 5. Indeed, most AS events identified in genes for which the e_T-ratio interquartile is higher in cultivated compartments corresponds to potentially 'functional' events (whereas AS events found in genes where e_T-ratio interquartile is lower in cultivated compartments are almost equally distributed between functional and non-functional).

To further confirm this hypothesis we looked closer at the genes showing the most extreme changes in e_T-ratio median value or interquartile. We found that the 'regulation of biological quality' GO annotation was over-represented among the genes for which the e_T-ratio interquartile in cultivated vs wild sorghum differs the most. Most genes showing the strongest AS e_T-ratio differences (outliers) are highly homologous to genes of other species shown to be involved in the genetic control of phenotypic traits related to the domestication syndrome. It could be worth to conduct a deeper functional analysis of the few remaining unannotated outlier genes. We are convinced that such AS e_T-ratio signatures could reveal domestication genes otherwise missed by more traditional methods of selection footprint detection or quantitative genetic approaches (QTL/GWAS).

Finally, in the same way that nucleotide diversity is a mutation reservoir on which natural selection acts, AS can be seen as a leverage on which selection may act too. It should also be kept in mind that AS is a mechanism which can be mobilized to respond to environmental stresses (recently reviewed in [23, 24, 25]). The loss of AS variability caused by domestication is contributing to the domestication load, and probably affects the adaptability potential of crops. This result also underlines the key importance of the conservation and management of the wild compartment to ensure its mobilization in the breeding process of cultivated genotypes.

Supporting information

S1 Table. Isoforms list. In this table are listed, for each isoform, 1) the locus ('gene_id'), 2) the cufflink_id, 3) the origin of the isoform (described in the publically available annotation or new identified isoform by 'cufflink'), 4) the predicted protein (when possible), 5) the length of the protein and 6) the identifier of the mRNA under which the sequence is named in the fasta file (S1 File). Note that for some alternative isoforms the start codon of the canonical isoform (given in the publically available annotation) is not present anymore in the alternative mRNA, making protein prediction hazardous, and is then noted 'start_codon_not_found'.
(XLSX)

S1 File. Fasta file containing the mRNA sequences of the 2794 isoforms.
(FASTA)

Acknowledgments

The authors are very grateful to the ARCAD project (Agropolis Resource Center for Crop Conservation, Adaptation and Diversity), a flagship project of the Agropolis Fondation which gave them the opportunity to study AS in the context of sorghum domestication. They also want to thank ICRISAT Gene bank for providing the cultivated and wild sorghum seeds with a IS prefix and the Center for Biological tropical Resources of Montpellier (CRB Tropicales de Montpellier) for providing the cultivated accessions with a SSM prefix.

Author Contributions

Conceptualization: Vincent Ranwez, Nathalie Chantret.

Data curation: Audrey Serra.

Formal analysis: Audrey Serra.

Funding acquisition: Nathalie Chantret.

Investigation: Vincent Ranwez, Nathalie Chantret.

Methodology: Vincent Ranwez, Audrey Serra, Nathalie Chantret.

Project administration: Vincent Ranwez, Nathalie Chantret.

Resources: David Pot.

Software: Vincent Ranwez, Audrey Serra.

Supervision: Vincent Ranwez, Nathalie Chantret.

Validation: Vincent Ranwez, David Pot, Nathalie Chantret.

Visualization: Vincent Ranwez, Nathalie Chantret.

Writing – original draft: Vincent Ranwez, Nathalie Chantret.

Writing – review & editing: Vincent Ranwez, David Pot, Nathalie Chantret.

References

1. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010; 463(7280):457–63. https://doi.org/10.1038/nature08909 PMID: 20110989.
2. Modrek B, Lee C. A genomic view of alternative splicing. Nature genetics. 2002; 30(1):13–9. https://doi.org/10.1038/ng0102-13 PMID: 11753382.
3. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012; 338(6114):1587–93. https://doi.org/10.1126/science.1230612 PMID: 23258890.
4. Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013; 14(3):153–65. https://doi.org/10.1038/nrm3525 PMID: 23385723.
5. Lopez-Diez R, Rastrojo A, Villate O, Aguado B. Complex tissue-specific patterns and distribution of multiple RAGE splice variants in different mammals. Genome Biol Evol. 2013; 5(12):2420–35. https://doi.org/10.1093/gbe/evt188 PMID: 24273313.
6. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009; 10(11):741–54. https://doi.org/10.1038/nrm2777 PMID: 19773805.
7. Caceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet. 2002; 18(4):186–93. PMID: 11932019.
8. Cieply B, Carstens RP. Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA. 2015; 6(3):311–26. https://doi.org/10.1002/wrna.1276 PMID: 25630614.
9. Kazan K. Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. Trends Plant Sci. 2003; 8(10):468–71. https://doi.org/10.1016/j.tplants.2003.09.001 PMID: 14557042.
10. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome research. 2010; 20(9):1238–49. https://doi.org/10.1101/gr.106120.110 PMID: 20627892.
11. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome research. 2012; 22(6):1184–95. https://doi.org/10.1101/gr.134106.111 PMID: 22391557.
12. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7(3):562–78. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036.

13. Walters B, Lum G, Sablok G, Min XJ. Genome-wide landscape of alternative splicing events in Brachypodium distachyon. DNA research: an international journal for rapid publication of reports on genes and genomes. 2013; 20(2):163–71. https://doi.org/10.1093/dnares/dss041 PMID: 23297300.

14. Potenza E, Racchi ML, Sterck L, Coller E, Asquini E, Tosatto SC, et al. Exploration of alternative splicing events in ten different grapevine cultivars. BMC genomics. 2015; 16(1):706. https://doi.org/10.1186/s12864-015-1922-5 PMID: 26380971.

15. Panahi B, Mohammadi SA, Ebrahimi Khaksefidi R, Fallah Mehrabadi J, Ebrahimie E. Genome-wide analysis of alternative splicing events in Hordeum vulgare: Highlighting retention of intron-based splicing and its possible function through network analysis. FEBS Lett. 2015; 589(23):3564–75. https://doi.org/10.1016/j.febslet.2015.09.023 PMID: 26454178.

16. Sun Y, Xiao H. Identification of alternative splicing events by RNA sequencing in early growth tomato fruits. BMC genomics. 2015; 16(1):948. https://doi.org/10.1186/s12864-015-2128-6 PMID: 26573826.

17. Panahi B, Abbaszadeh B, Taghizadeghan M, Ebrahimie E. Genome-wide survey of Alternative Splicing in Sorghum Bicolor. Physiol Mol Biol Plants. 2014; 20(3):323–9. https://doi.org/10.1007/s12298-014-0245-3 PMID: 25049459.

18. Chuang TJ, Yang MY, Lin CC, Hsieh PH, Hung LY. Comparative genomics of grass EST libraries reveals previously uncharacterized splicing events in crop plants. BMC Plant Biol. 2015; 15:39. https://doi.org/10.1186/s12870-015-0431-7 PMID: 25652661.

19. Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, et al. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol Biol Evol. 2012; 29(9):2265–83. https://doi.org/10.1093/molbev/mss100 PMID: 22446687.

20. Severing EI, van Dijk AD, Stiekema WJ, van Ham RC. Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC genomics. 2009; 10:154. https://doi.org/10.1186/1471-2164-10-154 PMID: 19358722.

21. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 2003; 100(1):189–92. https://doi.org/10.1073/pnas.0136770100 PMID: 12502788.

22. McGlincy NJ, Smith CWJ. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends in Biochemical Sciences. 2008; 33(8):385–93. https://doi.org/10.1016/j.tibs.2008.06.001 PMID: 18621535

23. Staiger D, Brown JW. Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell. 2013; 25(10):3640–56. https://doi.org/10.1105/tpc.113.113803 PMID: 24179132.

24. Filichkin S, Priest HD, Megraw M, Mockler TC. Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. Current opinion in plant biology. 2015; 24:125–35. https://doi.org/10.1016/j.pbi.2015.02.008 PMID: 25835141.

25. Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. BMC genomics. 2014; 15:431. https://doi.org/10.1186/1471-2164-15-431 PMID: 24897929.

26. Feng J, Li J, Gao Z, Lu Y, Yu J, Zheng Q, et al. SKIP Confers Osmotic Tolerance during Salt Stress by Controlling Alternative Gene Splicing in Arabidopsis. Mol Plant. 2015; 8(7):1038–52. https://doi.org/10.1016/j.molp.2015.01.011 PMID: 25617718.

27. Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, et al. Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. Plant Physiol. 2016; 170(1):586–99. https://doi.org/10.1104/pp.15.01267 PMID: 26582726.

28. Yang S, Tang F, Zhu H. Alternative splicing in plant immunity. Int J Mol Sci. 2014; 15(6):10424–45. https://doi.org/10.3390/ijms150610424 PMID: 24918296.

29. Meyer RS, DuVal AE, Jensen HR. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. 2012; 196(1):29–48. https://doi.org/10.1111/j.1469-8137.2012.04253.x PMID: 22889076.

30. Glemin S, Bataillon T. A comparative view of the evolution of grasses under domestication. New Phytol. 2009; 183(2):273–90. https://doi.org/10.1111/j.1469-8137.2009.02884.x PMID: 19515223.

31. Olsen KM, Wendel JF. Crop plants as models for understanding plant adaptation and diversification. Front Plant Sci. 2013; 4:290. https://doi.org/10.3389/fpls.2013.00290 PMID: 23914199.

32. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nature genetics. 2012; 44(7):808–11. https://doi.org/10.1038/ng.2309 PMID: 22660546.

33. Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, et al. Reshaping of the maize transcriptome by domestication. Proc Natl Acad Sci U S A. 2012; 109(29):11878–83. https://doi.org/10.1073/pnas.1201961109 PMID: 22753482.

34. Bao Y, Hu G, Flagel LE, Salmon A, Bezanilla M, Paterson AH, et al. Parallel up-regulation of the profilin gene family following independent domestication of diploid and allopolyploid cotton (Gossypium). Proc Natl Acad Sci U S A. 2011; 108(52):21152–7. https://doi.org/10.1073/pnas.1115926109 PMID: 22160709.

35. Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, et al. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci U S A. 2013; 110(28):E2655–62. https://doi.org/10.1073/pnas.1309606110 PMID: 23803858.

36. Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, et al. Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. Plant Cell. 2014; 26(5):1901–12. https://doi.org/10.1105/tpc.114.124040 PMID: 24850850.

37. Zou H, Tzarfati R, Hubner S, Krugman T, Fahima T, Abbo S, et al. Transcriptome profiling of wheat glumes in wild emmer, hulled landraces and modern cultivars. BMC genomics. 2015; 16:777. https://doi.org/10.1186/s12864-015-1996-0 PMID: 26462652.

38. Rapp RA, Haigler CH, Flagel L, Hovav RH, Udall JA, Wendel JF. Gene expression in developing fibres of Upland cotton (Gossypium hirsutum L.) was massively altered by domestication. BMC Biol. 2010; 8:139. https://doi.org/10.1186/1741-7007-8-139 PMID: 21078138.

39. Huang J, Gao Y, Jia H, Liu L, Zhang D, Zhang Z. Comparative transcriptomics uncovers alternative splicing changes and signatures of selection from maize improvement. BMC genomics. 2015; 16:363. https://doi.org/10.1186/s12864-015-1582-5 PMID: 25952680.

40. Frere CH, Prentis PJ, Gilding EK, Mudge AM, Cruickshank A, Godwin ID. Lack of low frequency variants masks patterns of non-neutral evolution following domestication. PLoS One. 2011; 6(8):e23041. https://doi.org/10.1371/journal.pone.0023041 PMID: 21853065.

41. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009; 457(7229):551–6. https://doi.org/10.1038/nature07723 PMID: 19189423.

42. Muraya MM, Mutegi E, Geiger HH, de Villiers SM, Sagnard F, Kanyenji BM, et al. Wild sorghum from different eco-geographic regions of Kenya display a mixed mating system. Theor Appl Genet. 2011; 122 (8):1631–9. https://doi.org/10.1007/s00122-011-1560-5 PMID: 21360157.

43. Mutegi E, Sagnard F, Labuschagne M, Herselman L, Semagn K, Deu M, et al. Local scale patterns of gene flow and genetic diversity in a crop-wild-weedy complex of sorghum (Sorghum bicolor (L.) Moench) under traditional agricultural field conditions in Kenya. CONSERVATION GENETICS. 2012; 13(4):1059–71. https://doi.org/10.1007/s10592-012-0353-y

44. Sagnard F, Deu M, Dembele D, Leblois R, Toure L, Diakite M, et al. Genetic diversity, structure, gene flow and evolutionary relationships within the Sorghum bicolor wild-weedy-crop complex in a western African region. Theor Appl Genet. 2011; 123(7):1231–46. https://doi.org/10.1007/s00122-011-1662-0 PMID: 21811819.

45. Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, et al. Ecogeographical distribution of wild, weedy and cultivated Sorghum bicolor (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genetic Resources and Crop Evolution. 2010; 57(2):243–53.

46. Lin Z, Li X, Shannon LM, Yeh CT, Wang ML, Bai G, et al. Parallel domestication of the Shattering1 genes in cereals. Nature genetics. 2012; 44(6):720–4. https://doi.org/10.1038/ng.2281 PMID: 22581231.

47. Barro-Kondombo C, Sagnard F, Chantereau J, Deu M, Vom Brocke K, Durand P, et al. Genetic structure among sorghum landraces as revealed by morphological variation and microsatellite markers in three agroclimatic regions of Burkina Faso. Theor Appl Genet. 2010; 120(8):1511–23. https://doi.org/10.1007/s00122-010-1272-2 PMID: 20180097.

48. Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquadro CF, et al. Challenges of detecting directional selection after a bottleneck: lessons from Sorghum bicolor. Genetics. 2006; 173(2):953–64. https://doi.org/10.1534/genetics.105.054312 PMID: 16547110.

49. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun. 2013; 4:2320. https://doi.org/10.1038/ncomms3320 PMID: 23982223.

50. Sarah G, Homa F, Pointet S, Contreras S, Sabot F, Nabholz B, et al. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour. 2016. https://doi.org/10.1111/1755-0998.12587 PMID: 27487989.

51. Clement Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. Evolutionary forces affecting synonymous variations in plant genomes. PLoS genetics. 2017; 13(5):e1006799. https://doi.org/10.1371/journal.pgen.1006799 PMID: 28531201.

52. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. 2011; 17:10–2.

53. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. https://doi.org/10.1093/bioinformatics/btp120 PMID: 19289445.

54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286.

55. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5):511–5. https://doi.org/10.1038/nbt.1621 PMID: 20436464.

56. Goldmark JP, Fazzio TG, Estep PW, Church GM, Tsukiyama T. The Isw2 chromatin remodeling complex represses early meiotic genes upon recruitment by Ume6p. Cell. 2000; 103(3):423–33. PMID: 11081629.

57. Dall'Osto L, Bressan M, Bassi R. Biogenesis of light harvesting proteins. Biochim Biophys Acta. 2015; 1847(9):861–71. https://doi.org/10.1016/j.bbabio.2015.02.009 PMID: 25687893.

58. Herritt M, Dhanapal AP, Fritschi FB. Identification of Genomic Loci Associated with the Photochemical Reflectance Index by Genome-Wide Association Study in Soybean. The Plant Genome. 2016; 9(21). https://doi.org/10.3835/plantgenome2015.08.0072 PMID: 27898827

59. Xia Y, Ning Z, Bai G, Li R, Yan G, Siddique KH, et al. Allelic variations of a light harvesting chlorophyll a/b-binding protein gene (Lhcb1) associated with agronomic traits in barley. PLoS One. 2012; 7(5): e37573. https://doi.org/10.1371/journal.pone.0037573 PMID: 22662173.

60. Li D, Zhu H, Liu K, Liu X, Leggewie G, Udvardi M, et al. Purple acid phosphatases of Arabidopsis thaliana. Comparative analysis and differential regulation by phosphate deprivation. J Biol Chem. 2002; 277 (31):27772–81. https://doi.org/10.1074/jbc.M204183200 PMID: 12021284.

61. Zhang Q, Wang C, Tian J, Li K, Shou H. Identification of rice purple acid phosphatases related to phosphate starvation signalling. Plant Biol (Stuttg). 2011; 13(1):7–15. https://doi.org/10.1111/j.1438-8677.2010.00346.x PMID: 21143719.

62. Zhang D, Song H, Cheng H, Hao D, Wang H, Kan G, et al. The acid phosphatase-encoding gene GmACP1 contributes to soybean tolerance to low-phosphorus stress. PLoS genetics. 2014; 10(1): e1004061. https://doi.org/10.1371/journal.pgen.1004061 PMID: 24391523.

63. Kane NC, Rieseberg LH. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, Helianthus annuus. Genetics. 2007; 175(4):1823–34. https://doi.org/10.1534/genetics.106.067728 PMID: 17237516.

64. Kajander T, Kellosalo J, Goldman A. Inorganic pyrophosphatases: one substrate, three mechanisms. FEBS Lett. 2013; 587(13):1863–9. https://doi.org/10.1016/j.febslet.2013.05.003 PMID: 23684653.

65. Khan MA, Olsen KM, Sovero V, Kushad MM, Korban SS. Fruit Quality Traits Have Played Critical Roles in Domestication of the Apple. The Plant Genome. 2014; 7(3). https://doi.org/10.3835/plantgenome2014.04.0018

66. Woods DP, Ream TS, Minevich G, Hobert O, Amasino RM. PHYTOCHROME C is an essential light receptor for photoperiodic flowering in the temperate grass, Brachypodium distachyon. Genetics. 2014; 198(1):397–408. https://doi.org/10.1534/genetics.114.166785 PMID: 25023399.

67. Saidou AA, Mariac C, Luong V, Pham JL, Bezancon G, Vigouroux Y. Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. Genetics. 2009; 182(3):899–910. https://doi.org/10.1534/genetics.109.102756 PMID: 19433627.

68. Lemmon ZH, Bukowski R, Sun Q, Doebley JF. The role of cis regulatory evolution in maize domestication. PLoS genetics. 2014; 10(11):e1004745. https://doi.org/10.1371/journal.pgen.1004745 PMID: 25375861.

69. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, et al. Genome-wide analysis of transcript isoform variation in humans. Nature genetics. 2008; 40(2):225–31. https://doi.org/10.1038/ng.2007.57 PMID: 18193047.

70. Hull J, Campino S, Rowlands K, Chan MS, Copley RR, Taylor MS, et al. Identification of common genetic variation that modulates alternative splicing. PLoS genetics. 2007; 3(6):e99. https://doi.org/10.1371/journal.pgen.0030099 PMID: 17571926.

71. Zhang W, Duan S, Bleibel WK, Wisel SA, Huang RS, Wu X, et al. Identification of common genetic variants that account for transcript isoform variation between human populations. Hum Genet. 2009; 125 (1):81–93. https://doi.org/10.1007/s00439-008-0601-x PMID: 19052777.

Evolutionary Dynamics of the Leucine-Rich Repeat Receptor-Like Kinase (LRR-RLK) Subfamily in Angiosperms[1][OPEN]

Iris Fischer*, Anne Diévart, Gaetan Droc, Jean-François Dufayard, and Nathalie Chantret*

Institut National de la Recherche Agronomique, Unité Mixte de Recherche Amélioration Génétique et Adaptation des Plantes Méditerranéennes et Tropicales, F–34060 Montpellier, France (I.F., N.C.); and Centre de Coopération Internationale en Recherche Agronomique Pour le Développement, Unité Mixte de Recherche AGAP, F–34398 Montpellier, France (A.D., G.D., J.-F.D.)

ORCID IDs: 0000-0002-5080-1223 (I.F.); 0000-0001-9460-4638 (A.D.); 0000-0003-1849-1269 (G.D.).

Gene duplications are an important factor in plant evolution, and lineage-specific expanded (LSE) genes are of particular interest. Receptor-like kinases expanded massively in land plants, and leucine-rich repeat receptor-like kinases (LRR-RLK) constitute the largest receptor-like kinases family. Based on the phylogeny of 7,554 LRR-RLK genes from 31 fully sequenced flowering plant genomes, the complex evolutionary dynamics of this family was characterized in depth. We studied the involvement of selection during the expansion of this family among angiosperms. LRR-RLK subgroups harbor extremely contrasting rates of duplication, retention, or loss, and LSE copies are predominantly found in subgroups involved in environmental interactions. Expansion rates also differ significantly depending on the time when rounds of expansion or loss occurred on the angiosperm phylogenetic tree. Finally, using a d_N/d_S-based test in a phylogenetic framework, we searched for selection footprints on LSE and single-copy LRR-RLK genes. Selective constraint appeared to be globally relaxed at LSE genes, and codons under positive selection were detected in 50% of them. Moreover, the leucine-rich repeat domains, and specifically four amino acids in them, were found to be the main targets of positive selection. Here, we provide an extensive overview of the expansion and evolution of this very large gene family.

Receptor-like kinases (RLKs) constitute one of the largest gene families in plants and expanded massively in land plants (Embryophyta; Lehti-Shiu et al., 2009, 2012). For plant RLK gene families, the functions of most members are often not known (especially in recently expanded families), but some described functions include innate immunity (Albert et al., 2010), pathogen response (Dodds and Rathjen, 2010), abiotic stress (Yang et al., 2010), development (De Smet et al., 2009), and sometimes multiple functions (Lehti-Shiu et al., 2012). The RLKs usually consist of three domains: an N-terminal extracellular domain, a transmembrane domain, and a C-terminal kinase domain (KD). In plants, the KD usually has a Ser/Thr specificity (Shiu and Bleecker, 2001), but Tyr-specific RLKs were also described (e.g. BRASSINOSTEROID INSENSITIVE1; Oh et al., 2009). Interestingly, it was estimated that approximately 20% of RLKs contain a catalytically inactive KD (e.g. STRUBBELIG and CORYNE; Chevalier et al., 2005; Castells and Casacuberta, 2007; Gish and Clark, 2011). In Arabidopsis (*Arabidopsis thaliana*), 44 RLK subgroups (SGs) were defined by inferring the phylogenetic relationships between the KDs (Shiu and Bleecker, 2001). Interestingly, different SGs show different duplication/retention rates (Lehti-Shiu et al., 2009). Specifically, RLKs involved in stress responses show a high number of tandemly duplicated genes whereas those involved in development do not (Shiu et al., 2004), which suggests that some RLK genes are important for the responses of land plants to a changing environment (Lehti-Shiu et al., 2012). There seem to be relatively few RLK pseudogenes compared with other large gene families, and copy retention was argued to be driven by both drift and selection (Zou et al., 2009; Lehti-Shiu et al., 2012). As most SGs are relatively old and RLK subfamilies expanded independently in several plant lineages, duplicate retention cannot be explained by drift alone, and natural selection is expected to be an important driving factor in RLK gene family retention (Lehti-Shiu et al., 2009).

[1] This work was supported by the German Research Foundation (grant no. FI 1984/1–1 to I.F.), by the Agropolis Resource Center for Crop Conservation, Adaptation, and Diversity project of the Agropolis Foundation (to I.F.), and by the Agence Nationale de la Recherche (grant no. ANR–08–GENM–021).

* Address correspondence to irisfischer402@gmail.com and chantret@supagro.inra.fr.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Nathalie Chantret (chantret@supagro.inra.fr).

I.F., A.D., and N.C. designed the study; G.D. performed the LRR-RLK extraction; J.-F.D. performed the phylogenetic clustering; I.F., A.D., and N.C. performed the data and statistical analyses; I.F. wrote the article with the help of A.D. and N.C.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.15.01470

Leucine-rich repeat-receptor-like kinases (LRR-RLKs), which contain up to 30 leucine-rich repeat (LRRs) in their extracellular domain, constitute the largest RLK family (Shiu and Bleecker, 2001). Based on the KD, 15 LRR-RLK SGs have been established in Arabidopsis (Shiu et al., 2004; Lehti-Shiu et al., 2009). So far, two major functions have been attributed to them: defense against pathogens and development (Tang et al., 2010b). LRR-RLKs involved in defense are predominantly found in lineage-specific expanded (LSE) gene clusters, whereas LRR-RLKs involved in development are mostly found in nonexpanded groups (Tang et al., 2010b). It was also discovered that the LRR domains are significantly less conserved than the remaining domains of the LRR-RLK genes (Tang et al., 2010b). In addition, a study of four plant genomes (Arabidopsis, grape [*Vitis vinifera*], poplar [*Populus trichocarpa*], and rice [*Oryza sativa*]) showed that LRR-RLK genes from LSE gene clusters show significantly more indications of positive selection or relaxed constraint than LRR-RLKs from nonexpanded groups (Tang et al., 2010b).

The genomes of flowering plants (angiosperms) have been shown to be highly dynamic compared with most other groups of land plants (Leitch and Leitch, 2012). This dynamic is mostly caused by the frequent multiplication of genetic material, followed by a complex pattern of differential losses (i.e. the fragmentation process) and chromosomal rearrangements (Langham et al., 2004; Leitch and Leitch, 2012). Most angiosperm genomes sequenced so far show evidence for at least one whole-genome multiplication event during their evolution (Jaillon et al., 2007; D'Hont et al., 2012; Tomato Genome Consortium, 2012). At a smaller scale, tandem and segmental duplications are also very common in angiosperms (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Rizzon et al., 2006). Although the most common fate of duplicated genes is to be progressively lost, in some cases they can be retained in the genome, and adaptive as well as nonadaptive scenarios have been discussed to play a role in this preservation process (for review, see Moore and Purugganan, 2005; Hahn, 2009; Innan, 2009; Innan and Kondrashov, 2010). Whole-genome sequences also revealed that the same gene may undergo several rounds of duplication and retention. These LSE genes were shown to evolve under positive selection more frequently than single-copy genes in angiosperms (Fischer et al., 2014). That study analyzed general trends over whole genomes. Here, we ask if, and to what extent, this trend is observable at LRR-RLK genes. As this gene family is very dynamic and large, and in accordance with the results of Tang et al. (2010b), we expect the effect of positive selection to be even more pronounced than in the whole-genome average.

We analyzed 33 Embryophyta genomes to investigate the evolutionary history of the LRR-RLK gene family in a phylogenetic framework. Twenty LRR-RLK SGs were identified, and from this data set, we deciphered the evolutionary dynamics of this family within angiosperms. The expansion/reduction rates were contrasted between SGs and species as well as in ancestral branches of the angiosperm phylogeny. We then focused on genes whose number increased dramatically in an SG- and/or species-specific manner (i.e. LSE genes). Those genes are likely to be involved in species-specific cellular processes or adaptive interactions and were used as a template to infer the potential occurrence of positive selection. This led to the identification of sites at which positive selection likely acted. We discuss our results in the light of angiosperm genome evolution and current knowledge of LRR-RLK functions. Positive selection footprints identified in LSE genes highlight the importance of combining evolutionary analysis and functional knowledge to guide further investigations.

RESULTS

We extracted genes containing both LRRs and a KD from 33 published embryophyte genomes. Here, we mostly describe the findings for the 31 angiosperm (eight monocot and 23 dicot) genomes we analyzed. The 7,554 LRR-RLK genes were classified in 20 SGs. This classification was inferred using distance-related methods, because the high number of sequences to be analyzed would imply excessive computation time for methods relying on maximum likelihood. Since we decided to study the evolutionary dynamic of the LRR-RLK gene family using SG classification as a starting point, we first wanted to verify that each SG was monophyletic. Ten subsets of about 750 sequences were created by picking one sequence out of 10 to infer a PHYML tree (data not shown). Analysis of the trees shows that most SGs (14) are monophyletic with strong branch support. On the other hand, for six SGs (SG_I, SG_III, SG_VI, SG_Xb, SG_XI, and SG_XV), the topology differs slightly between trees: in at least five trees out of 10, either the SG appears to be paraphyletic or few sequences are placed outside the main monophyletic clade with low branch support. As we could not confirm that these SGs are monophyletic, they were tagged with an asterisk throughout this article.

Next, we determined the number of ancestral genes present in the last common ancestor of angiosperms (LCAA) using a tree reconciliation approach (see "Materials and Methods"). In short, tree reconciliation compares each SG-specific LRR-RLK gene tree with the species tree to infer gene duplications and losses. Note that since only LRR-RLKs with at least one complete LRR were considered, some of the inferred gene losses might correspond to RLKs without, or with degenerated, LRRs. Using this method, we predicted the number of LRR-RLK genes in the LCAA to be 150. All SGs were present in the LCAA, but the number of genes between SGs was highly variable (Table I). SG_III* and SG_XI* show the highest number of ancestral genes, with 32 and 29 genes, respectively. The lowest numbers of ancestral genes are recorded for SG_VIIb, SG_Xa, SG_XIIIa, and SG_XIIIb, which only possessed two genes, and SG_XIV, which only contained one. These results show that, already in

Table I. *Total number of LRR-RLKs in our angiosperm data set, number of ancestral genes in the LCAA, and median global expansion rate for each SG among the 31 species*

SG	Total No. of Genes	No. of Ancestral Genes	Median Global Expansion Rate
I*	482	7	2.00
II	349	9	1.22
III*	1,400	32	1.22
IV	131	3	1.33
V	263	5	1.80
VI*	324	10	1.00
VIIa	157	3	1.67
VIIb	84	2	1.50
VIII-1	216	5	1.40
VIII-2	355	8	1.25
IX	193	3	1.67
Xa	143	2	2.00
Xb*	367	9	1.11
XI*	1,177	29	1.28
XIIa	1,126	9	3.00
XIIb	423	4	2.00
XIIIa	84	2	1.50
XIIIb	77	2	1.00
XIV	84	1	3.00
XV*	119	5	0.80
Total	7,554	150	

the LCAA, which lived approximately 150 million years ago (Supplemental Table S1), some SGs were more prone to retain copies than others. We wanted to determine if this ancestral pattern was preserved during the course of angiosperm evolution and if different SGs expanded or contracted compared with the LCAA.

Expansion Rates of LRR-RLK Genes Differ between Subgroups and Species

To gain a more comprehensive understanding of LRR-RLK evolution, we first looked at SG-specific expansion rates in two complementary ways. First, we calculated the global SG expansion rate (the ratio of contemporary LRR-RLK genes per species in one SG divided by the ancestral number) for each SG (Fig. 1). Second, we inferred the branch-specific expansion rate of each SG on the phylogenetic tree of the 31 angiosperm species. We did this by automatically computing the ratio of descendant LRR-RLKs divided by the ancestral number of LRR-RLKs at every node (see "Materials and Methods"; Fig. 2). Looking at the global SG expansion, we found that SG_Xa, SG_XIIa, SG_XIIb, and SG_XIV expanded more than 2-fold on average, and SG_I* and SG_IX expanded around 2-fold (Fig. 1; Supplemental Table S2). Interestingly, SG_XIIa already had a moderately high ancestral gene number (nine) and, therefore, seems to be generally prone to high retention rates. Indeed, SG_XIIa was subject to repeated rounds of major expansion events (i.e. expansion greater than 2-fold) during its evolutionary history (e.g. in Poaceae, the *Solanum* ancestor, Malvaceae, and the Arabidopsis ancestor) but also species-specific expansions, e.g. in THECC, GOSRA, ARALY, SCHPA, MALDO, LOTJA, POPTR, and JATCU (Fig. 2; for five-digit species codes, see Table II). On the other hand, SG_I* and SG_XIIb had a medium number of copies in the ancestral genome (seven and four, respectively) but the pattern of expansion is quite different when analyzed in detail (Fig. 2). For SG_I*, the expansion rate is mostly due to ancestral expansion events rather than species-specific ones. For example, the high number of copies in ARATH and EUTSA (Fig. 1) is not due to expansions specific to these species but rather an expansion in Brassicaceae. Subsequently, copies were lost in the other species of this family analyzed here (ARALY, SCHPA, BRARA) but retained in ARATH and EUTSA (Fig. 2). Species-specific expansions can also be observed in SG_I*, mostly in PRUPE and POPTR. For SG_XIIb, on the other hand, the high expansion rate is mostly due to recent species-specific expansions in PHODC, MUSAC, VITVI, GOSRA, MALDO, POPTR, and JATCU. But one major ancestral expansion can be observed in Rosids.

SG_IX, SG_Xa, and SG_XIV had only a few copies in the LCAA (three, two, and one, respectively), and all show a relatively high global expansion rate (Fig. 1). For these SGs also, a contrasted branch-specific expansion pattern can be observed (Fig. 2). SG_Xa went through relatively few major expansions: one can be detected in the dicot ancestor and a species-specific one in POPTR. Likewise, SG_IX shows only one ancestral expansion in Malvaceae but more species-specific expansions in PHODC, MUSAC, MALDO, and GLYMA. Finally, SG_XIV went through several rounds of ancestral (monocots, dicots, Malvaceae, and Brassicaceae) as well as species-specific (PHODC, MALDO, and POPTR) expansion. The other SGs show a moderate expansion rate (1.3–1.75) or no expansion at all (Fig. 1; Supplemental Table S2). SG_XV* is the only SG for which the number of copies was decreasing on average compared with the LCAA genome (0.77). It is important to note that the LCAA ancestral gene number could have been overestimated slightly for those SGs without a confirmed monophyletic origin (denoted by asterisks), resulting in an underestimation of global expansion rate. However, we recalculated the global expansion rates for each of those SGs using the largest subset of sequences that always include a stable monophyletic clade. The obtained global expansion rate differed only slightly from the ones presented here (data not shown), and the conclusions drawn remain unchanged.

Because some species underwent whole-genome duplication (WGD) or whole-genome triplication (WGT) relatively recently compared with others (Table III), we determined species-specific patterns of LRR-RLK expansions and determined if those patterns are consistent with the recent history of the species. Therefore, we computed the global species expansion rate (the ratio of LRR-RLK genes per SG in one species divided by the ancestral number) for each of the 31 angiosperm species. As expected, the global expansion rate differs significantly between species (Fig. 3; Supplemental Table S2). Compared with the LCAA (150 genes), the number of LRR-RLK genes did not

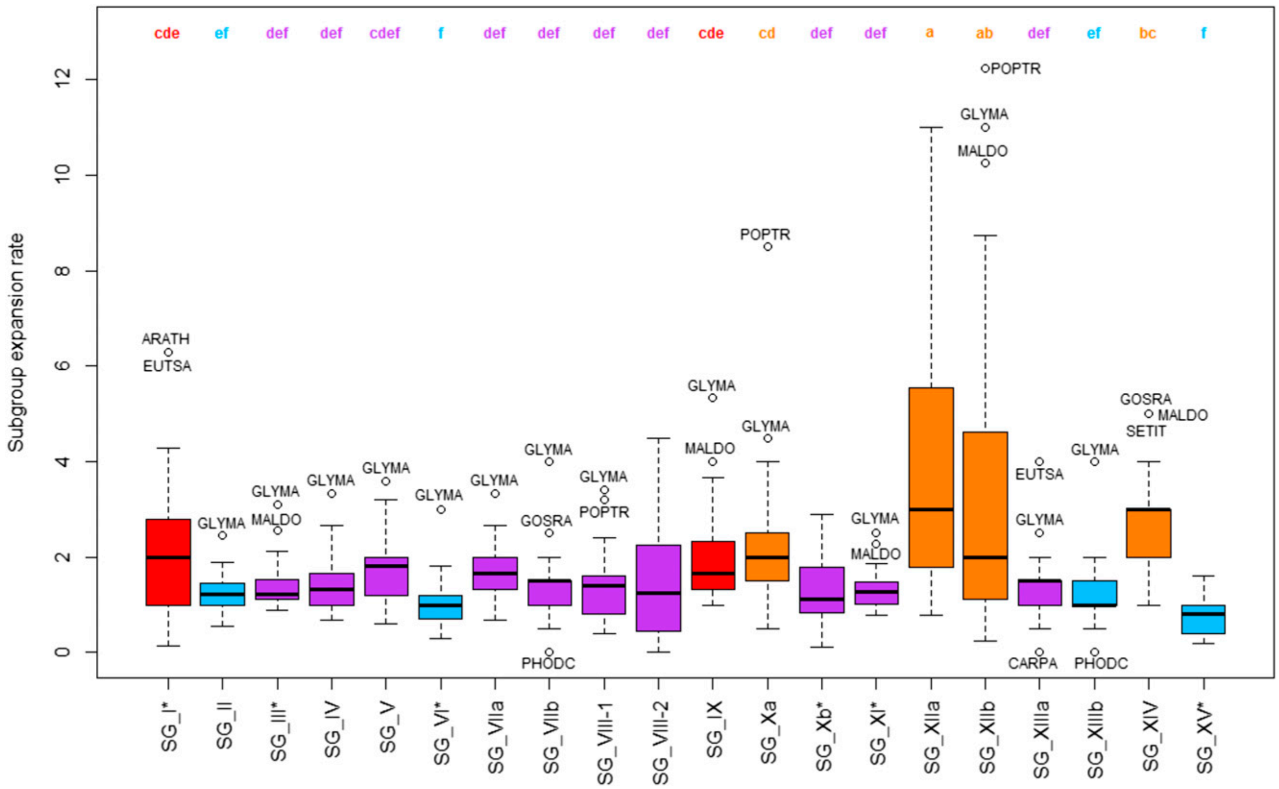


Figure 1. Global expansion rate in each SG, which is the total number of genes in each species divided by the ancestral number (Table I). An ANOVA test showed that the expansion rate differs significantly between SGs ($P < 2\text{e-}16$). Therefore, we performed a TukeyHSD test to determine which SGs exactly show a significant difference between each other and grouped those SGs by significance level (a–e). Letters above each box plot indicate the TukeyHSD significance group (Supplemental Table S2). The significance groups are color coded according to the mean expansion rate: orange, greater than 2.25-fold expansion; red, 1.75- to 2.25-fold expansion; purple, 1.3- to 1.75-fold expansion; and blue, 0.75- to 1.3-fold expansion (i.e. no expansion). The outlier species are labeled for each SG. For species identifiers, see Table II.

decrease for most species except for LOTJA (114) and CARPA (127). This indicates that, on average, LRR-RLK genes are more prone to retention than loss. Some species, however, did not significantly expand their average number of LRR-RLK genes compared with the common ancestor: PHODC (158), CUCME (149), CUCSA (180), SCHPA (194), BRARA (185), MEDTR (183), and RICCO (182). LRR-RLK genes expanded more than 2-fold in GLYMA (477), MALDO (441), POPTR (400), and GOSRA (372) and around 2-fold in MUSAC (280), MAIZE (241), SETIT (301), ORYSJ (317), ORYSI (301), SOLTU (254), PRUPE (260), MANES (238), and EUTSA (240). The remaining species show a moderate expansion rate (1.4–1.75): CACJA (222), THECC (238), JATCU (208), ARATH (222), SOLLC (232), SORBI (225), BRADI (225), VITVI (193), and ARALY (195). As expected, the four species with the highest global expansion rate (GLYMA, MALDO, POPTR, and GOSRA) are recent polyploids in which most SGs have expanded (Fig. 2). However, some SGs expanded more than 2-fold, indicating that small-scale duplication events have occurred in addition to polyploidy. In POPTR, for instance, the global expansion rates of SG_Xa and SG_XIIb are more than 8-fold (Fig. 3), and a strong branch-specific expansion rate is detected on the terminal POPTR branch (3.25 for SG_Xa and 5.4 for SG_XIIb; Fig. 2). Surprisingly, SG_VIIa and SG_VIIb show a high branch-specific expansion rate in POPTR (4 and 3, respectively), which is not reflected in the global expansion rate in this species (Fig. 3). This is due to the fact that SG_VIIa and SG_VIIb went through strong reduction in Malpighiales (0.33) and fabids (0.5), respectively. Thus, the cumulative effect of successive reductions and expansions is not evident in the global expansion rate. These contrasted evolutionary dynamics can also be observed in MALDO. A global expansion of SG_IX was not detected because of the strong reduction in Amygdaloideae. To summarize, these data can be integrated into the species phylogeny to draw an image of the complex evolutionary dynamics of the LRR-RLK gene family through time (Fig. 4).

Different Patterns of Lineage-Specific Expansion in LRR-RLK Subgroups

Given the differences of LRR-RLK expansion rates between species, we wanted to identify cases of LSE (i.e. cases where a high duplication/retention rate is specific

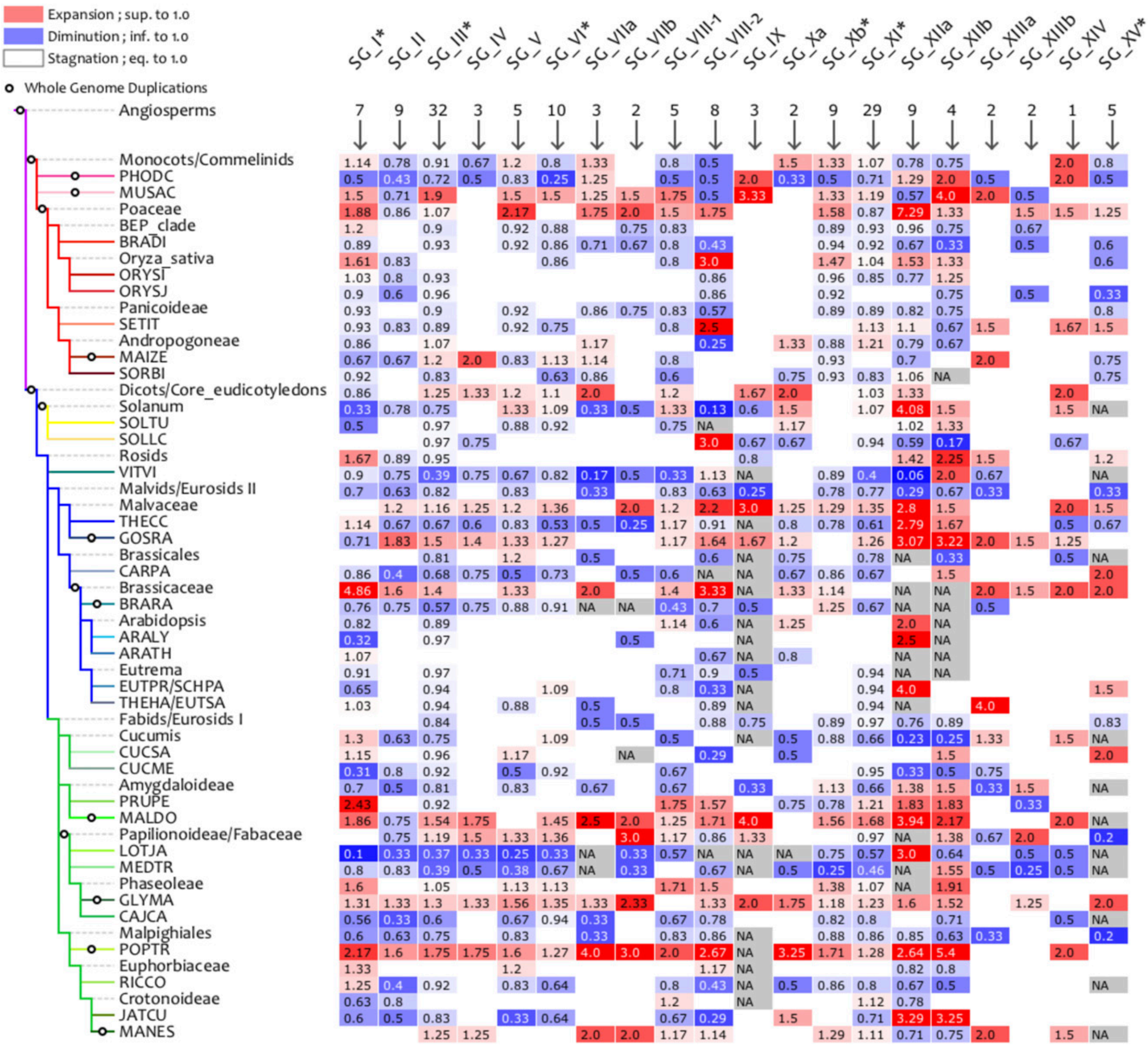


Figure 2. Branch-specific expansion/diminution of LRR-RLK genes for every SG on every branch in the phylogenetic tree. The tree on the left displays all the nodes and branches, and polyploidy events are marked with dots. Every line gives the expansion rate where the current (descendant) node is compared with the previous (ascendant) node. Red boxes indicate expansion, blue boxes indicate diminution, and blank boxes indicate stagnation. For example: SG_I* has the same number of copies in monocots compared with the ascendant node (angiosperms) indicated by a blank box. In PHODC, a diminution occurred compared with the ascendant node (monocots) indicated by a blue box. In MUSAC, an expansion occurred compared with the ascendant node (monocots) indicated by a red box, and so on.

to one species). Using a tree reconciliation approach (see "Materials and Methods"), we built a data set consisting of ultraparalog (UP; related only by duplication) clusters that represents the LSE events and a superortholog (SO; related only by speciation) reference gene set. We only considered clusters containing five or more sequences. After cleaning, our final data set comprised 75 UP and 189 SO clusters containing 796 and 1,970 sequences, respectively (Table IV). The median number of sequences in the UP clusters is not significantly different from the median number in the SO clusters (eight in both cases; Supplemental Fig. S1). For UP clusters, however, the alignments are significantly longer (Mann-Whitney test, $P < 0.001$), with a median of 3,237 bp compared with 2,841 bp for SO clusters. One possible explanation for this could be that UP clusters are more dynamic and might contain more LRRs. PRANK, the alignment algorithm we used, introduces gaps instead of aligning ambiguous sites and, therefore, produces longer alignments when sequences are divergent. However, this phenomenon does not

Table II. *Five-digit code for each species*

Species Name	Common Name	Five-Digit Code
Phoenix dactylifera	Date palm	PHODC
Musa acuminata	Banana	MUSAC
Brachypodium distachyon	Purple false brome	BRADI
Oryza sativa ssp. *japonica*	Asian rice	ORYSJ
Oryza sativa ssp. *indica*	Indian rice	ORYSI
Setaria italica	Foxtail millet	SETIT
Zea mays	Maize	MAIZE
Sorghum bicolor	Milo	SORBI
Solanum tuberosum	Potato	SOLTU
Solanum lycopersicum	Tomato	SOLLC
Vitis vinifera	Common grape vine	VITVI
Theobroma cacao	Cacao tree	THECC
Gossypium raimondii	Cotton progenitor	GOSRA
Carica papaya	Papaya	CARPA
Arabidopsis thaliana	Thale cress	ARATH
Arabidopsis lyrata	Outcrossing Arabidopsis relative	ARALY
Brassica rapa	Turnip	BRARA
Schrenkiella parvula	A saltwater cress	SCHPA
Eutrema salsugineum	A saltwater cress	EUTSA
Cucumis sativus	Cucumber	CUCSA
Cucumis melo	Melon	CUCME
Prunus persica	Peach	PRUPE
Malus × domestica	Apple	MALDO
Lotus japonicus		LOTJA
Medicago truncatula	Barrel medic	MEDTR
Glycine max	Soybean	GLYMA
Cajanus cajan	Pigeon pea	CAJCA
Populus trichocarpa	Black cottonwood	POPTR
Ricinus communis	Castor oil plant	RICCO
Jatropha curcas	Barbados nut	JATCU
Manihot esculenta	Cassava	MANES
Selaginella moellendorffii	A spikemoss	SELML
Physcomitrella patens	A moss	PHYPA

influence the outcome of further tests for positive selection using codeml (Yang, 2007).

We then wanted to determine which SGs are represented in the SO and UP data sets. Unsurprisingly, all SGs were present in SO clusters (Fig. 5). This was expected, as all SGs were already present in the LCAA and remained stable or expanded (except SG_XV*). In general, the frequency of SO clusters (and sequences) for each SG reflects the number of copies in the LCAA (Table I; Fig. 5). On the other hand, only 11 of the 20 SGs were represented in UP clusters (SG_I*, SG_III*, SG_VI*, SG_VIII-2, SG_IX, SG_Xa, SG_Xb*, SG_XI*, SG_XIIa, SG_XIIb, and SG_XIIIa), and these SGs harbor a total of 837 sequences. SG_I*, SG_VIII-2, SG_XIIa, and SG_XIIb are clearly overrepresented, which is in accordance with their expansion pattern. Other expanded SGs, however, have only a low number of UP clusters or, in the case of SG_IV, no UP clusters at all. Therefore, it seems that recently duplicated genes are more prone to be retained in some SGs.

Differences of Selective Constraint between Subgroups, Domains, and Amino Acids

To provide further insight into the LRR-RLK gene family evolution, we wanted to determine under which kind of selective pressures the LRR-RLK genes evolved. We focused on the data set described above (i.e. LSE and orthologous genes). We inferred the d_N/d_S ratio (or ω, i.e. the ratio of nonsynonymous to synonymous substitution rates) at codons of the alignments and branches of the phylogeny of the UP and SO clusters. An $\omega = 1$ indicates neutral evolution/relaxed constraint, an $\omega < 1$ indicates purifying selection, and an $\omega > 1$ can indicate positive selection. We used mapNH (Dutheil et al., 2012; Romiguier et al., 2012) to compute the ω for each branch. mapNH ran for 71 UP and 176 SO clusters containing 1,246 and 2,960 branches, respectively (Table IV). We first wanted to test for relaxation of selective constraint in UP and SO clusters and looked for branches with $\omega > 1$. We found 6.04% of UP branches but only 0.49% of SO branches to have an $\omega > 1$. The mean ω for branches with $\omega > 1$ is significantly larger in UP clusters (1.45) compared with SO clusters (1.13; $P = 0.004$). The same is true for branches with $\omega < 1$, where ω is significantly larger in UP clusters (0.48) compared with SO clusters (0.24; $P < 0.001$). Overall, the mean ω is significantly larger for branches from UP clusters (0.54) than for SO clusters (0.24; $P < 0.001$; Table IV; Supplemental Fig. S2).

We found 38 out of 75 UP clusters (50.67%) containing codons under positive selection (for details, see Supplemental Table S3) after manual curation but only six out of 186 SO clusters (3.23%). Additionally, codons under positive selection found in UP clusters are not distributed evenly over domains (Fig. 6). To account for the differences in domain size, a hit frequency (i.e. the number of sites under positive selection we found relative to all sites possible for each domain) was calculated (see "Materials and Methods"). The domain showing the highest hit frequency is the LRR domain, followed by the Cys pairs and their flanking regions (Fig. 6A). Hits in both domains are distributed over all SGs and species tested. The KD and its surrounding domains contain very few codons under positive selection. Domains classified as other combine domains important for the function of the LRR-RLK genes but vary between SGs. For example, SG_I* (Fig. 6B) contains a malectin domain. All hits classified as other here fall in the malectin-like domain of a POPTR SG_I* cluster.

Finally, we wanted to investigate whether some amino acids in the LRR are more frequently targeted by positive selection. The LRR typically contains 24 amino acids and sometimes islands between them (Fig. 6C). Four amino acids were predominantly subject to positive selection: 6, 8, 10, and 11, which all lie in the LRR-characteristic LXXLXLXX β-sheet/β-turn structure.

DISCUSSION

We studied the SG- and species-specific expansion dynamics in LRR-RLK genes from 31 angiosperm genomes in a phylogenetic framework. We also analyzed

Table III. *Estimated times of polyploidy events and corresponding references for Figure 4*

Event	Name	Reference	Age
			million years
1	Seed plant tetraploidy	Jiao et al. (2011)	350–330
2	Angiosperm tetraploidy	Jiao et al. (2011)	230–190
3	Monocot tetraploidy	Tang et al. (2010a)	130
4	Date palm WGD	D'Hont et al. (2012)	75–65 (?)
5	Banana gamma	D'Hont et al. (2012)	100
6	Banana beta	D'Hont et al. (2012)	65
7	Banana alpha	D'Hont et al. (2012)	65
8	Grass tetraploidy B (sigma)	D'Hont et al. (2012)	123–109
9	Grass tetraploidy (rho)	Paterson et al. (2004)	70
10	Maize tetraploidy	Schnable et al. (2011)	12–5
11	Eudicot hexaploidy (Arabidopsis gamma)	Jaillon et al. (2007); Cenci et al. (2010); Wang et al. (2012)	150–120
12	*Solanum* hexaploidy	Tomato Genome Consortium (2012)	91–52
13	Papiloniod tetraploidy	Pfeil et al. (2005)	55–54
14	Soybean tetraploidy	Pfeil et al. (2005)	15–13
15	Apple tetraploidy	Velasco et al. (2010); Verde et al. (2013)	45–30
16	Poplar tetraploidy	Tuskan et al. (2006)	65–60
17	Arabidopsis beta	Fawcett et al. (2009)	70–40
18	Arabidopsis alpha	Barker et al. (2009)	23
19	*Brassica* hexaploidy	Wang et al. (2011)	9–5
20	Cotton WGD	Wang et al. (2012)	20–13
21	Cassava WGD	Mühlhausen and Kollmar (2013)	? (after Crotonoideae split)

the lineage-specifically expanded genes in this family to determine to what extent positive selection occurred on them using a d_N/d_S-based test. We found differences in expansion patterns depending on SGs and species but only a few SGs that were subject to LSE. A significantly higher proportion of LSE LRR-RLK genes was affected by positive selection compared with single-copy genes, and the LRR domain (specifically four amino acids within this domain) was targeted by positive selection. In the following, we will discuss our findings in more detail.

Subgroup- and Species-Specific Expansions

We observed significant variations in the global expansion rates between LRR-RLK SGs. These are due to a complex history of expansion-retention-loss cycles that are specific to each SG. The phylogenetic approach allowed us to determine the relative importance of ancestral versus recent species-specific expansions for each SG and to characterize precisely the loss/retention dynamics during the evolutionary history of the studied species (summarized in Fig. 4). For example, SG_III* and SG_XI* had a high copy number of LRR-RLKs in the LCAA and kept a stable copy number over the last 150 million years. On the other hand, SG_I*, SG_XIIa, and SG_XIIb, which had a moderate copy number in the LCAA, keep expanding. Some functions have been described for genes of these SGs, mainly in Arabidopsis (Supplemental Table S4). For SG_III* and SG_XI*, mostly genes involved in development are described. The high numbers of ancestral genes in these two SGs combined with their size stability during angiosperm evolution may be interpreted as an early high level of diversification/specialization of these genes that are needed to orchestrate common developmental features. This hypothesis can be reinforced by the high number of superorthologous genes in these SGs. For SG_I* and SG_XIIa, on the other hand, mostly genes involved in responses to biotic stress are described at present. These observations confirm that different expansion/retention patterns appear to be related to gene function, although one has to keep in mind that functions have only been assigned to a few LRR-RLK genes. Three SGs (SG_IX, SG_Xa, and SG_XIV) expanded compared with their very low ancestral number (one to three), leading to a high total expansion rate. As it has been postulated that duplications are the raw material for adaptation (Nei and Rooney, 2005; Fischer et al., 2014), the evolution of those SGs was likely driven by adaptation, to varying degrees in different angiosperm species, depending on the environment they evolved in. The known functions are both related to responses to biotic or abiotic stress and development. Because so far our knowledge of LRR-RLK functions is limited and mostly restricted to Arabidopsis, further studies are needed to make more reliable statements on the link between function and expansion/retention dynamics in different SGs.

Next, we wanted to ascertain species-specific expansions of LRR-RLK genes and how they are related to the recent history of the species in our study. Whole-genome multiplication has been argued to be a major force in the diversification of angiosperms (Soltis et al., 2009; Soltis and Burleigh, 2009; Renny-Byfield and Wendel, 2014). All angiosperms share two ancient WGDs (Jiao et al., 2011). Likewise, all monocots share a WGD approximately 130 million years ago (Tang et al.,

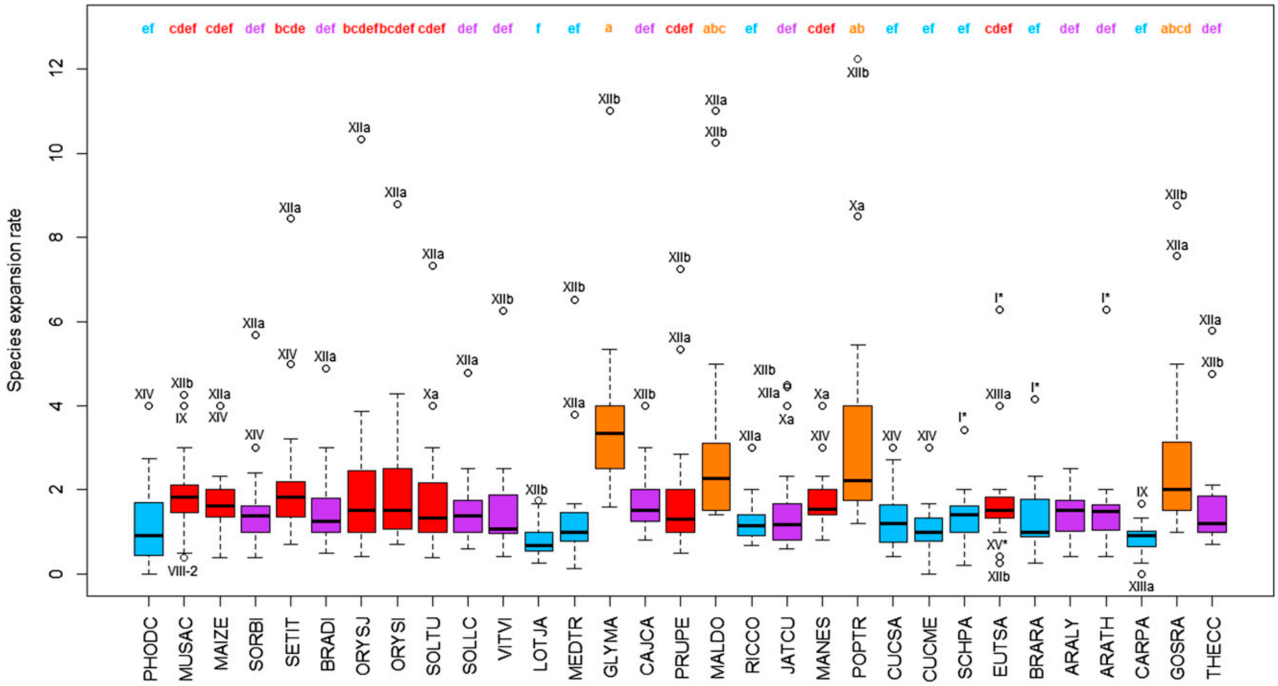


Figure 3. Global expansion rate in each species, which is the total number of genes in each species divided by the ancestral number (Table I). An ANOVA test showed that the expansion rate differs significantly between species ($P < 2e\text{-}16$). Therefore, we performed a TukeyHSD test to determine which species exactly show a significant difference between each other and grouped those species by significance level (a–e). Letters above each box plot indicate the TukeyHSD significance group (Supplemental Table S2). The significance groups are color coded according to the mean expansion rate: orange, greater than 2.25-fold expansion; red, 1.75- to 2.25-fold expansion; purple, 1.4- to 1.75-fold expansion; and blue, 0.8- to 1.4-fold expansion (i.e. no expansion). The outlier SGs are labeled for each species. For species identifiers, see Table II.

2010a), and most dicots (eudicots) share a WGT around the same time (Jaillon et al., 2007; Wang et al., 2012), but more recent WGDs and WGTs occurred in many angiosperm species (Fig. 4; Table III). The link between WGD/WGTs and the number of LRR-RLK genes is not straightforward. We found that in soybean (*Glycine max*), *Gossypium raimondii*, and apple (*Malus* × *domestica*), which were subject to relatively recent WGDs (15–13, 17–13, and 45–30 million years ago, respectively; Pfeil et al., 2005; Velasco et al., 2010; Wang et al., 2012), the number of LRR-RLK genes expanded more than 2-fold compared with the LCAA. These results are in accordance with what was already described for these species. Indeed, it was found that soybean contains a very large number of retained genes from this WGD (Cannon et al., 2015). Additionally, recent studies on large gene families in *G. raimondii* indicate that their copy number is driven either by retention after the last WGD (e.g. NAC transcription factors; Shang et al., 2013) or by a combination of segmental duplications (SDs) and tandem duplications (TDs; e.g. WRKY transcription factors; Dou et al., 2014). For apple (most recent WDG after the divergence for peach [*Prunus persica*] according to Verde et al. [2013]), a recent study on nucleotide-binding site LRR genes showed that they also stem mostly from SDs and TDs (Arya et al., 2014).

More contrasting results are observed in the Brassicaceae, where two WGDs occurred (Barker et al., 2009; Fawcett et al., 2009). Most SGs expand their number of genes on this ancestral branch, but the species belonging to this clade mostly retain or lose genes on average (Figs. 2 and 4). The only exception concerns *Eutrema salsugineum* (an Arabidopsis relative), which is the only species with a greater than 2-fold average expansion rate. The global expansion rate in *E. salsugineum* is mostly due to two SGs (SG_I* and SG_XIIIa). In the original genome study (Wu et al., 2012), the authors found that genes from the category response to stimulus (response to salt stress, osmotic stress, water deprivation, abscisic acid stimulus, and hypoxia) are significantly overrepresented in *E. salsugineum* compared with Arabidopsis. This overrepresentation is described as mostly caused by SDs and TDs (Wu et al., 2012), in accordance with what we observed in SG_XIIIa. This could be of functional importance to this halophile plant.

Finally, of all species analyzed here, maize (*Zea mays*) and *Brassica rapa* (and maybe *Manihot esculenta*) show the most recent cases of WGD/WGT (12–5 and 9–5 million years ago, respectively; Schnable et al., 2011; Wang et al., 2011), yet their expansion rates are moderate. This is further evidence for the dynamic nature of angiosperm genomes that has been discussed before

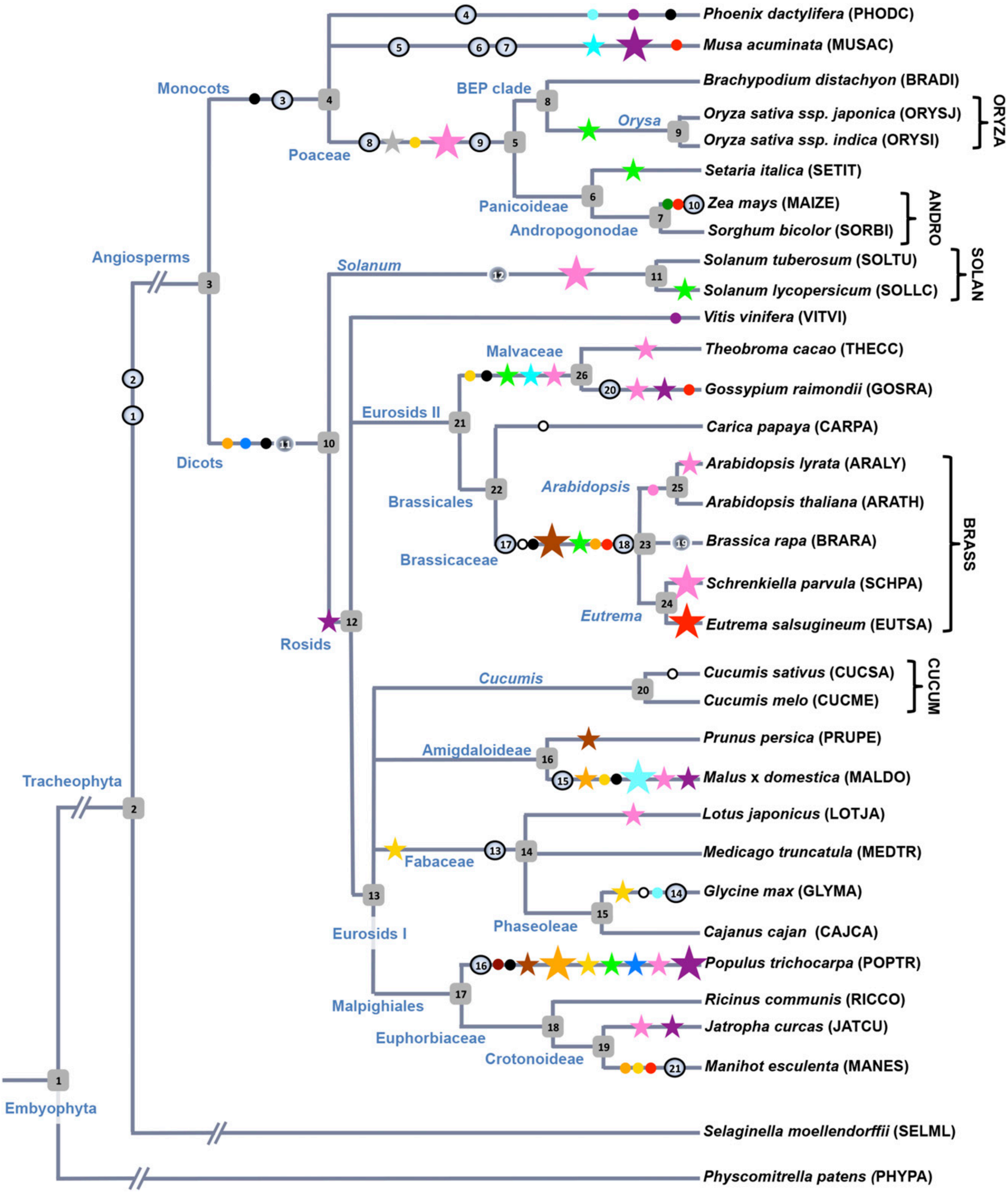


Figure 4. Phylogenetic tree of the 33 species studied here. Five-digit species identifiers are given in parentheses next to the species names. Species that diverged less than 15 million years ago were merged for the LSE analysis (see "Materials and Methods"): ANDRO, ORYZA, SOLAN, CUCUM, and BRASS. Polyploidy events and their estimated ages are indicated on the tree: circles on the branches represent WGD, and dark circles represent WGT. The numbers in the circles refer to details on the polyploidization events given in Table I. Species divergence and their estimated age are indicated by gray squares on the nodes. The numbers in the squares refer to details on the divergence times given in Supplemental Table S1. Dots and asterisks on the branches indicate SG expansions: dots, 2-fold; small asterisks, between 2- and 4-fold; and large asterisks, equal to or more than

Table IV. Details of the LSE and mapNH analyses for UP and SO clusters

Parameter	UP	SO
Total No. of clusters	75	189
Clusters for final mapNH analysis	71	176
Median cluster size (first; third Qu)	8 (6; 12)	8 (6; 14)
Minimum; maximum cluster size	5; 38	5; 25
Median alignment length (first; third Qu)	3,237 (2,952; 3,574)	2,841 (2,034; 3,192)
Minimum; maximum alignment length	1,749; 8,691	861; 6,216
Branches analyzed/total No. of branches	1,193/1,246	2,860/2,960
Clusters with branches $\omega > 1$ (%)	25 (35.21)	10 (5.68)
Branches with $\omega < 1$ (%)	1,121 (93.96)	2,846 (99.51)
Mean ω for less than one branch ± SD	0.48 ± 0.17	0.24 ± 0.12
Branches with $\omega > 1.0$ (%)	72 (6.04)	14 (0.49)
Mean ω for more than one branch ± SD	1.45 ± 0.51	1.13 ± 0.14
Mean ω ± SD	0.54 ± 0.31	0.24 ± 0.13

(Leitch and Leitch, 2012; Fischer et al., 2014). After a WGD event, genomes tend to return to the diploid (or previous) state by losing redundant duplicated genes (fractionation process), although the gene loss is biased (Bowers et al., 2003; Schnable et al., 2009). Which genes are lost or retained depends strongly on their function (De Smet et al., 2013). However, it has been shown that genes involved in stress responses are mostly created by TD rather than WGD (Hanada et al., 2008). Indeed, it was hypothesized before that RLK genes involved in stress responses mostly duplicate by TD (Shiu et al., 2004). Here, we provide a detailed representation of expansion-retention-loss dynamics of the whole LRR-RLK gene family in 31 angiosperm species (Fig. 4). Each new genome sequenced will improve the accuracy of the expansion-retention-loss event predictions and will help in identifying new elements that can be useful for future functional analysis and/or linked to adaptive traits.

Studying Selection Pressures in a Large and Dynamic Gene Family

As described above, the composition of LRR-RLKs in each of the 31 studied angiosperm species results from a complex dynamic of species- and SG-specific expansion/loss events. To further investigate the potential role of this family in plant adaptation, we analyzed the selective pressures to which the LRR-RLKs were subject. Such an analysis cannot be considered for the phylogeny of the entire gene family because of the high number of sequences and the high sequence divergence (the phylogeny on which we divided the SGs was inferred on the conserved KD only). We then chose to focus on two specific cases: (1) LSE as a specific case of duplication/retention, and (2) a subset of strictly orthologous genes. Indeed, LSE has been shown to fuel adaptation in angiosperms (Fischer et al., 2014), and we wanted to test the prevalence of this mode of duplication in our large data set. Therefore, we evaluated the extent to which LRR-RLK genes were subject to LSE and how positive selection acted on those genes. As a reference, we chose the strictly orthologous subset. This approach allows the interpretation of LSE evolution compared with the general LRR-RLK selective background (Fischer et al., 2014).

The power of this phylogenetic approach relies on the number of species analyzed, and we profit from an ever-increasing number of sequenced plant genomes. Another important requirement for this approach is the quality of sequencing and annotation, especially for a large gene family, as sequencing errors and misannotations can lead to false positives when testing for positive selection (Han et al., 2013). We profit from a recently developed pipeline designed to automatically perform different steps of the analysis (Fischer et al., 2014). This allowed us to quickly incorporate sequenced genomes of choice, and future studies can easily expand this analysis as new reliable data become available. Finally, we set great value on manually verifying the data throughout the process, from the identification of the LRR-RLKs to the inference of positive selection. Although this is tedious work for such a large data set, it is important nevertheless. As we recently showed, approximately 50% automatically reported instances of positive selection turned out to be false positives after manual curation (Fischer et al., 2014).

We found that all SGs are represented in the single-copy reference set, with an overrepresentation of SG_III* and SG_XI*. This is in accordance with the fact that these two SGs had the highest number of copies in the genome of the LCAA and did not expand significantly since (see above). In general, the frequency of clusters from the single-copy gene set (and sequences) for each SG reflects the number of copies in the LCAA (Table I; Fig. 5). On the other hand, only 11 of the 20 SGs

Figure 4. (*Continued.*)
4-fold. SGs are as follows: SG I* (brown), SG_IV (dark green), SG_V (gray), SG_VIIa (orange), SG_VIIb (yellow), SG_VIII-1 (dark brown), SG_VIII-2 (green), SG_IX (light blue), SG_Xa (dark blue), SG_XIIa (pink), SG_XIIb (purple), SG_XIIIa (red), SG_XIV (black), and SG_XV* (white). The asterisks and dots do not indicate the exact age.

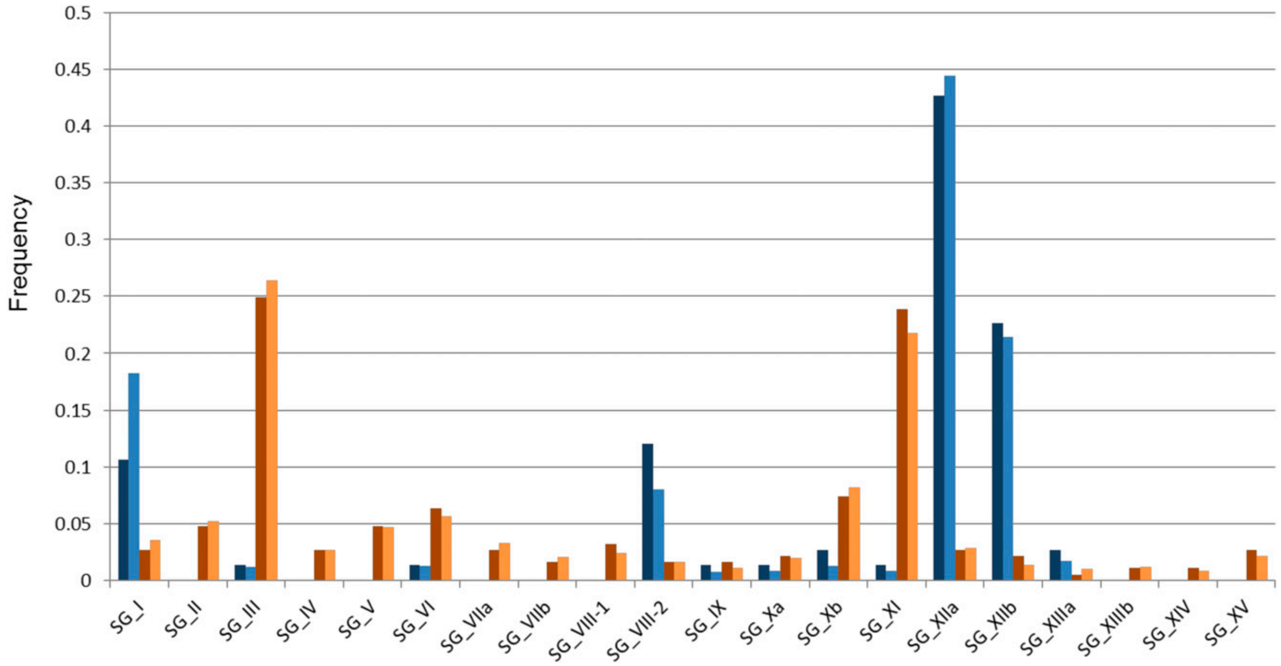


Figure 5. Distribution of UP and SO clusters and sequences across all SGs. The frequency of all extracted UP (dark blue) and SO (dark orange) clusters for each SG, and the frequency of all extracted UP (light blue) and SO sequences (light orange) for each SG, are shown.

were represented in the LSE data set. This is mainly because the majority of expansions are rather old in these SGs, whereas they happened relatively recently in SG_I*, SG_VIII-2, SG_XIIa, and SG_XIIb (see above). Fourteen species (or clades) are represented in the LSE data set: MUSAC (two UP clusters), SETIT (one), ORYZA (10), VITVI (three), SOLAN (six), MEDTR (three), GLYMA (two), PRUPE (six), MALDO (11), POPTR (eight), BRASS (11), GOSRA (five), THECC (two), and PHYPA (five). Again, not every species is affected to the same extent, but this does not necessarily reflect recent WGD/WGT. Additionally, LSE can also arise from SD and TD, the frequency of which is not uniform within or between genomes. Our results indicate that different species are more likely to retain recently duplicated genes than others. This, in turn, might reflect on their recent evolution or domestication, which should be examined in more detail in future studies.

When focusing on the study of selective pressures, we first looked at ω at the branches of the LSE and single-copy gene clusters and found that selective constraint was relaxed in the LSE data set. This outcome was expected, as it was shown previously that LSE genes evolve more relaxed constraint than single-copy genes in angiosperms (Fischer et al., 2014). This study, however, looked at whole angiosperm genomes, but a similar pattern has already been demonstrated in other large gene families (Johnson and Thomas, 2007; Xue et al., 2012; Yang et al., 2013a, 2013b) and in LRR-RLK genes in particular (Tang et al., 2010b). Previous studies on that subject only had a limited data set (four angiosperm species; Tang et al., 2010b). Here, we demonstrate that this is still true when a larger and more representative sample of angiosperms is considered.

Next, we wanted to identify codons that evolved under positive selection in the LSE and the single-copy data sets. A recent study on gene families in the whole genomes of 10 angiosperms found that 5.4% of LSE genes contained codons showing positive selection footprints (Fischer et al., 2014). Here, we ask if and to what extent this is also true for the large and dynamic LRR-RLK gene family. We discovered that for LSE LRR-RLK genes, the rate of codons under selection is almost 10-fold higher (50.67%) than the genome average. In addition, we found more than 3% of single-copy genes containing codons under selection, whereas Fischer et al. (2014) described no case of positive selection at the single-copy gene clusters in their study. Together with the high rate of branches with $\omega > 1$ in LSE gene clusters (6.04%, compared with 0.49% for single-copy genes), this indicates that LRR-RLK genes are more prone to evolve under positive selection than the average for angiosperm gene families. As might be expected, all UP clusters with codons under positive selection come from the four overrepresented SGs: SG_I* (one UP cluster), SG_VIII-2 (three), SG_XIIa (24), and SG_XIIb (10). The single-copy gene clusters with codons under selection come from six SGs: SG_III, SG_VIIa, SG_Xa, SG_Xb, SG_XIIa, and SG_XIIb. Therefore, recent expansion and retention affect only a few SGs, but in those SGs positive selection plays an important role. For SG_XIIa, positive selection has been inferred previously for genes involved in environmental interactions: *Xa21*, which confers resistance to the bacterial blight disease, was found to have

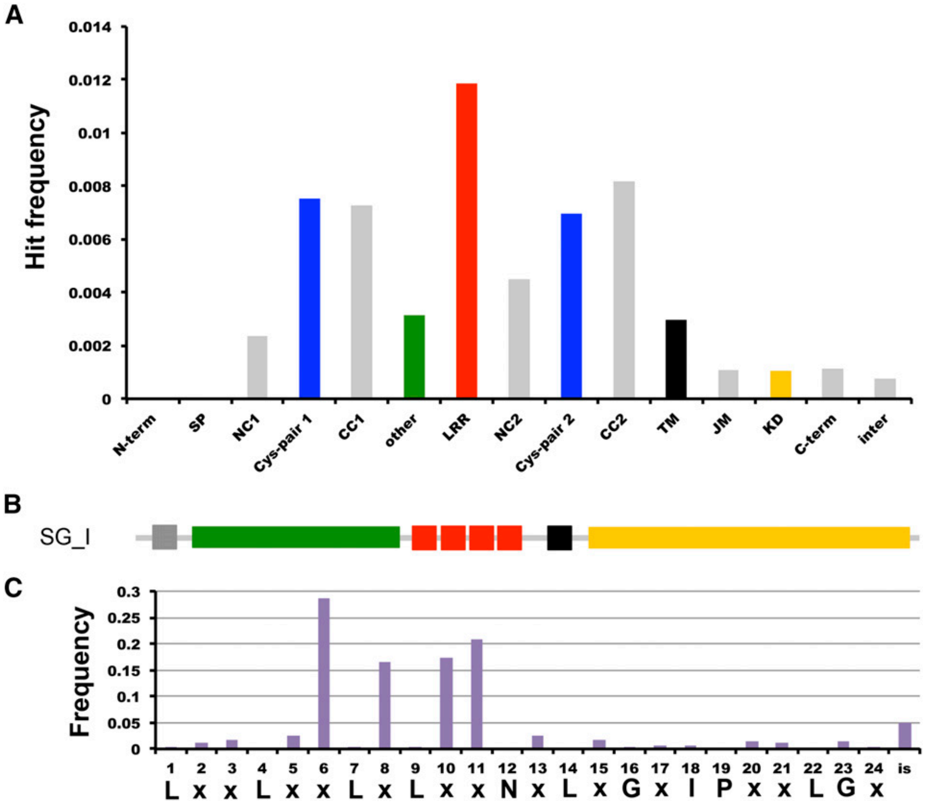


Figure 6. A, Hit frequency (i.e. frequency of codons under selection versus the total number of sites) for each domain of the LRR-RLK genes. The absence/presence and size of the domains vary between SGs, for details, see text. N-term, N-Terminal end; SP (dark gray), signal peptide; NC1, N-terminal end of Cys-pair 1; Cys-pair 1 (blue), first Cys pair; CC1, C-terminal end of Cys-pair 1; other (green), other domains; NC2, N-terminal end of Cys-pair 2; Cys-pair 2 (blue), second Cys pair; CC2, C-terminal end of Cys-pair 2; TM (black), transmembrane domain; JM, juxtamembrane domain; C-term, C-terminal end; inter, other interdomain regions. B, Schematic structure of the LRR-RLK genes, here with SG_I* gene structure as an example. C, Frequency of amino acids in the LRR domain under positive selection. L, Leu; x, variable; N, Asn; G, Gly; I, Ile; P, Pro; is, island between LRRs.

evolved under positive selection in rice (Wang et al., 1998; Tan et al., 2011); and FLS2, involved also in responses to biotic stress, shows a signature of rapid fixation of an adaptive allele in Arabidopsis (Vetter et al., 2012). Future studies on smaller subsets of SGs will surely cast further light on selection patterns in LRR-RLK genes. Only 11 species (or clades) are represented in the LSE data set with codons under positive selection: SETIT (one UP cluster), ORYZA (two), SOLAN (four), MEDTR (two), GLYMA (two), PRUPE (three), MALDO (eight), POPTR (seven), BRASS (two), GOSRA (five), and THECC (two). Not every species is affected to the same extent by positive selection, and again, future studies might bring more details concerning the evolutionary history of specific species and SGs to light.

In addition, we found that not every domain of the LRR-RLK genes was similarly affected by positive selection. Most codons under selection fall in the LRR domain. This outcome might be expected, as LRRs are very dynamic and plasticity in this region provides plants with a broad tool set to face environmental challenges and, therefore, undergoes positive selection frequently (Zhang et al., 2006; Tang et al., 2010b). Only very few codons under positive selection were found in the KD and its surrounding regions. This result is consistent with the fact that the KD is very conserved among species and SGs and evolved mostly under purifying selection (Shiu et al., 2004; Tang et al., 2010b). A more surprising result was the identification of a significant number of positively selected sites in the malectin-like domain of a poplar SG_I* cluster. So far, the function of extracellular malectin-like domains of RLKs is not well understood (Lindner et al., 2012). However, a malectin-like domain-containing SG_I* LRR-RLK has been described to confer susceptibility to a downy mildew pathogen in Arabidopsis and to have similarities to symbiosis RLKs, which are important for the regulation of bacterial symbiont accommodation (Markmann et al., 2008; Hok et al., 2011). Therefore, our results suggest that it could be interesting to further investigate the function and evolutionary history of this SG_I* domain, particularly in poplar. Another unexpected finding was the frequent occurrence of positive selection at the Cys pairs and flanking regions that are involved in folding and/or the binding to other proteins. To what extent the function of LRR-RLKs is affected by mutations in the Cys pair regions depends on the function of the gene (Song et al., 2010; Sun et al., 2012), and it would be interesting to study this in more detail in the future.

Finally, we took a closer look at the amino acids in the LRR primarily affected by positive selection. Only four, out of the 24 amino acids an LRR typically contains, were predominantly and strongly subject to positive selection. These variable amino acids lie in the unconserved part of the LRR-characteristic LXXLXLXX β-sheet/β-turn structure, which is involved in protein-protein interactions (Jones and Jones, 1997; Enkhbayar et al., 2004). Specifically, solvent-exposed residues were targeted by positive selection (Parniske et al., 1997;

Wang et al., 1998). Further investigation of the functional consequences of these nucleotide variations need to be done to confirm their adaptive potential, but our findings align very well with the current understanding of LRR ligand binding. Taken together, our results could be very useful for further functional investigations of LRR-RLK genes in different species.

CONCLUSION

We studied LRR-RLK genes from 33 land plant species to investigate SG- and species-specific expansion of these genes, the extent to which they were subject to LSE, and the role that positive selection played in the evolution of this large gene family. We described that some SGs are more prone to expansion/retention than others and that the expansions occurred at different times in the evolution of LRR-RLK genes. This fine-scale analysis of the dynamic allowed us to identify branches and species for which a higher than average retention rate could indicate a potential adaptive event for some SGs. We also described that only a few SGs show patterns of recent LSE and that, at those genes, selective constraint is relaxed. More than 50% of the LSE genes contain codons that show evidence for positive selection, which is almost 10-fold the frequency described previously for gene families in angiosperms (Fischer et al., 2014). Finally, we found that, across the LRR-RLK genes, the LRR domain and specifically four amino acids responsible for ligand interaction are most frequently subject to selection.

MATERIALS AND METHODS

Studied Genomes

We analyzed 31 angiosperm genomes (eight monocot [sub]species and 23 dicot species; Table II): *Phoenix dactylifera* (Al-Dous et al., 2011), *Musa acuminata* (D'Hont et al., 2012), *Oryza sativa* ssp. *japonica* (International Rice Genome Sequencing Project, 2005), *Oryza sativa* ssp. *indica* (Yu et al., 2002), *Brachypodium distachyon* (International Brachypodium Initiative, 2010), *Zea mays* (Schnable et al., 2009), *Sorghum bicolor* (Paterson et al., 2009), *Setaria italica* (Zhang et al., 2012), *Solanum tuberosum* (Xu et al., 2011), *Solanum lycopersicum* (Tomato Genome Consortium, 2012), *Vitis vinifera* (Jaillon et al., 2007), *Lotus japonicus* (Sato et al., 2008), *Cajanus cajan* (Varshney et al., 2012), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Arabidopsis lyrata* (Hu et al., 2011), *Schrenkiella parvula* (a synonym is *Eutrema parvula*; we used the nomenclature from Oh et al. [2014]; Dassanayake et al., 2011), *Eutrema salsugineum* (a synonym is *Thellungiella halophila*; we chose the nomenclature according to Phytozome [http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Esalsugineum]; Wu et al., 2012), *Brassica rapa* (Wang et al., 2011), *Populus trichocarpa* (Tuskan et al., 2006), *Glycine max* (Schmutz et al., 2010), *Medicago truncatula* (Young et al., 2011), *Prunus persica* (Ahmad et al., 2011), *Malus* × *domestica* (Velasco et al., 2010), *Ricinus communis* (Chan et al., 2010), *Jatropha curcas* (Sato et al., 2011), *Manihot esculenta* (Prochnik et al., 2012), *Cucumis sativus* (Huang et al., 2009), *Cucumis melo* (Garcia-Mas et al., 2012), *Carica papaya* (Ming et al., 2008), *Gossypium raimondii* (Wang et al., 2012), and *Theobroma cacao* (Argout et al., 2011). We also extracted LRR-RLKs from the moss *Physcomitrella patens* (Rensing et al., 2008) and the spikemoss *Selaginella moellendorffii* (Banks et al., 2011). Throughout this article, we refer to the species using five-digit identifiers, which can be found in Table II. Altogether, we analyzed 33 genomes from 39 proteomes (we used several annotation versions of the Arabidopsis and rice genomes). Details on which genome versions we used can be found in Supplemental Table S5. The phylogeny of those species is provided in Figure 4.

LRR-RLK Extraction, Clustering, Phylogeny, and Identification of Gain/Loss Events

We used the hmmsearch program (Eddy, 2009) to extract peptide sequences containing both intact (i.e. nondegenerated) LRR(s) and a KD from the proteomes as described previously (Diévart et al., 2011). We classified SGs using the KD by a global phylogenetic analysis (the tree can be found at http://phylogeny.southgreen.fr/kinase/index.php; Global Analysis). First, sequences were aligned using MAFFT (Katoh et al., 2005) with a progressive strategy. Second, the alignments were cleaned using trimAl (Capella-Gutiérrez et al., 2009) with settings to remove every site with more than 20% gaps or with a similarity score lower than 0.001. Third, a similarity matrix was computed by ProtDist (Felsenstein, 1993) using a JTT model. Fourth, a global distance phylogeny was inferred using FastME (Desper and Gascuel, 2006) with default settings and SPR movements to optimize the tree topology. Fifth, SGs were defined manually in the global phylogeny using the Arabidopsis genes as reference, which led us to 20 SGs in contrast to the 15 described previously (Shiu et al., 2004; Lehti-Shiu et al., 2009).

More accurate phylogenies were then inferred for each of the 20 SGs. The KDs of the sequences attributed to each SG were realigned using MAFFT with an iterative strategy (maximum of 100 iterations). Alignments were cleaned using trimAl with settings to only remove sites with more than 80% gaps. Then, maximum likelihood phylogenies were inferred by PhyML 3.0 (Guindon et al., 2010) using an LG+gamma model and the best of NNI and SPR topology optimization. Statistical branch support was computed using the aLRT/SH-like strategy (Guindon et al., 2010). This left us with 20 phylogenies, one for each SG (all phylogenies are available at http://phylogeny.southgreen.fr/kinase/index.php; SG_I–SG_XV).

Each of the 20 phylogenetic trees has been reconciled with the species tree using RAP-Green (Dufayard et al., 2005; https://github.com/SouthGreenPlatform/rap-green). By comparing the gene tree with the species tree, this analysis allows us to root phylogenetic trees and to infer duplication and loss events (Dufayard et al., 2005). We tested this approach of rooting (by minimizing the number of inferred duplications and losses) and compared it with rooting with outgroups (data not shown). The two methods provided very close root locations that did not change the overall conclusions. Using this RAP-Green tree reconciliation approach (for parameters, the maximum support for reduction is 0.95), we inferred the number of duplications and losses at each node of the species tree. Briefly, each duplication and loss increases and decreases, respectively, by one the number of copies in the common ancestor of the taxonomic group analyzed.

We determined the global SG- and species-specific expansion rates by computing the number of LRR-RLK genes in one SG divided by the ancestral number and the number of LRR-RLK genes in one species divided by the ancestral number, respectively. An ANOVA showed that the expansion rate differed significantly between the SGs/species ($P < 2\text{e-}16$ in both cases). We used the TukeyHSD test of the agricolae package (http://cran.r-project.org/web/packages/agricolae/index.html) in R (R Development Core Team, 2012) to further explore which groups of SGs/species differ from each other. This test compares the range of sample means and defines an honestly significance difference value, which is the minimum distance between groups to be considered statistically significant. In short, TukeyHSD is a posthoc test that groups subsets by significance levels after ANOVA showed significant differences between subsets.

LSE Data Set and Testing for Positive Selection

Testing for adaptation can be done by comparing positive (Darwinian) selection footprints in lineages with recently and specifically duplicated genes to reference lineages containing only single-copy genes. One way to infer positive selection is by analyzing nucleotide substitution data at the codon level in a phylogenetic framework. As nucleotide substitutions can be either nonsynonymous (i.e. protein changing, thereby potentially impacting the fitness) or synonymous (i.e. not protein changing, thereby theoretically without consequences for the fitness; Lawrie et al., 2013), the nonsynonymous/synonymous substitution rate ratio, denoted as d_N/d_S or ω, can be used to infer the direction and strength of natural selection. An $\omega < 1$ indicates purifying selection, and the closer ω is to 0, the stronger purifying selection is acting. Under neutral evolution, $\omega = 1$. An $\omega > 1$ indicates that positive selection is acting.

We identified UP clusters (related only by duplication) using a tree reconciliation approach (Dufayard et al., 2005). Those represent our LSE gene set. As a single-copy gene reference, we chose an SO gene set (related only by speciation). We chose clusters with a minimum of five sequences. To address the question of whether positive selection is more frequent after LSE events, we compared the results obtained on UPs with those obtained on SO gene sets. Species that diverged less than 15 million years ago were merged for the LSE detection (Fig. 4) in order not to overly reduce the UP data set and to not induce bias due to very recent speciation events: ANDRO (ZEAMA and SORBI), ORYZA (ORYSJ and ORYSI), SOLAN (SOLLC and SOLTU), CUCUM (CUCSA and CUCME), and BRASS (ARATH, ARALY, BRARA, SCHPA, and EUTSA). We then applied the pipeline developed by Fischer et al. (2014) to the extracted UP and SO clusters. In short, the pipeline consists of the following steps. (1) The clusters were aligned using PRANK$_{+F}$ with codon option (Löytynoja and Goldman, 2005). The alignments were cleaned by GUIDANCE (Penn et al., 2010) with the default sequence quality cutoff and a column cutoff of 0.97 to remove problematic sequences and unreliable sites from the alignments. We used PRANK and GUIDANCE here because previous benchmarks (Fletcher and Yang, 2010; Jordan and Goldman, 2012) showed that these programs lead to a minimum of false positives when inferring positive selection using codeml. The cleaned alignments can be retrieved at http://phylogeny.southgreen.fr/kinase/alignments.php (manually curated alignments for positive selection analysis). (2) We relied on the EggLib package (De Mita and Siol, 2012) to infer the maximum likelihood phylogeny at the nucleotide level for every alignment using PhyML 3.0 (Guindon et al., 2010) under the GTR substitution model. (3) We ran the codeml site model implemented in the PAML4 software (Yang, 2007) to infer positive selection on codons under several substitution models. In clusters identified to have evolved under positive selection, Bayes empirical Bayes was used to calculate the posterior probabilities at each codon and detect those under positive selection (i.e. those with a posterior probability of $\omega > 1$ strictly above 95%). All alignments detected to be under positive selection at the codon level were curated manually for potential alignment errors. Details of all codons showing a signal of positive selection using codeml can be found in Supplemental Table S3. (4) We used mapNH (Dutheil et al., 2012; Romiguier et al., 2012) to infer ω at the branch level.

In order to analyze the distribution of positively selected sites among domains, we calculated a hit frequency that computes the number of sites under positive selection found in each domain relative to all sites possible. All possible sites for each domain were calculated as follows. First, we extracted the size of each domain of every SG. If SGs were subdivided further, we took the average size of each domain. Second, we multiplied the size of each domain by the number of UP clusters we found for each SG. For example: the LRR of SG_I* contains an average of 77 sites, and we found eight UP clusters for SG_I*. Therefore, the total number of possible LRR sites for SG_I* is 77 × 8 = 616 sites. Third, we added up the sites for each domain for all SGs.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Summary of UP and SO cluster size and length.

Supplemental Figure S2. ω distribution of branches of UP and SO clusters.

Supplemental Table S1. Estimated divergence times and corresponding references for Figure 4.

Supplemental Table S2. Results of the TukeyHSD test.

Supplemental Table S3. Details of all codons showing a signal of positive selection using codeml.

Supplemental Table S4. Arabidopsis LRR-RLK gene classification according to The Arabidopsis Information Resource.

Supplemental Table S5. List of genomes used here, with name, link, and version of the genome fasta file.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments.

Received September 16, 2015; accepted January 14, 2016; published January 15, 2016.

LITERATURE CITED

Ahmad R, Parfitt DE, Fass J, Ogundiwin E, Dhingra A, Gradziel TM, Lin D, Joshi NA, Martinez-Garcia PJ, Crisosto CH (2011) Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. BMC Genomics **12:** 569

Albert M, Jehle AK, Mueller K, Eisele C, Lipschis M, Felix G (2010) *Arabidopsis thaliana* pattern recognition receptors for bacterial elongation factor Tu and flagellin can be combined to form functional chimeric receptors. J Biol Chem **285:** 19035–19042

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al (2011) *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotechnol **29:** 521–527

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al (2011) The genome of *Theobroma cacao*. Nat Genet **43:** 101–108

Arya P, Kumar G, Acharya V, Singh AK (2014) Genome-wide identification and expression analysis of NBS-encoding genes in *Malus × domestica* and expansion of NBS genes family in Rosaceae. PLoS ONE **9:** e107987

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al (2011) The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. Science **332:** 960–963

Barker MS, Vogel H, Schranz ME (2009) Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol Evol **1:** 391–399

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Rolf M, et al (2015) Multiple polyploidy events in the early radiation of nodulating and non-nodulating legumes. Mol Biol Evol **32:** 193–210

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25:** 1972–1973

Castells E, Casacuberta JM (2007) Signalling through kinase-defective domains: the prevalence of atypical receptor-like kinases in plants. J Exp Bot **58:** 3503–3511

Cenci A, Combes MC, Lashermes P (2010) Comparative sequence analyses indicate that *Coffea* (asterids) and *Vitis* (rosids) derive from the same paleo-hexaploid ancestral genome. Mol Genet Genomics **283:** 493–501

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al (2010) Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol **28:** 951–956

Chevalier D, Batoux M, Fulton L, Pfister K, Yadav RK, Schellenberg M, Schneitz K (2005) *STRUBBELIG* defines a receptor kinase-mediated signaling pathway regulating organ development in *Arabidopsis*. Proc Natl Acad Sci USA **102:** 9074–9079

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert HJ, et al (2011) The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet **43:** 913–918

De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet **13:** 27

De Smet I, Voss U, Jürgens G, Beeckman T (2009) Receptor-like kinases shape the plant. Nat Cell Biol **11:** 1166–1173

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA **110:** 2898–2903

Desper R, Gascuel O (2006) Getting a tree fast: neighbor joining, FastME, and distance-based methods. Curr Protoc Bioinformatics **Chapter 6:** 6.3.1–6.3.28

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature **488:** 213–217

Diévart A, Gilbert N, Droc G, Attard A, Gourgues M, Guiderdoni E, Périn C (2011) Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. BMC Evol Biol **11:** 367

Dodds PN, Rathjen JP (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet **11:** 539–548

Dou L, Zhang X, Pang C, Song M, Wei H, Fan S, Yu S (2014) Genome-wide analysis of the WRKY gene family in cotton. Mol Genet Genomics **289:** 1103–1121

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics **21:** 2596–2603

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B (2012) Efficient selection of branch-specific models of sequence evolution. Mol Biol Evol **29:** 1861–1874

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform **23:** 205–211

Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N (2004) Structural principles of leucine-rich repeat (LRR) proteins. Proteins Struct Funct Bioinf **54:** 394–403

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci USA **106:** 5737–5742

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle

Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard JF, Chantret N (2014) Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol **14:** 151

Fletcher W, Yang Z (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol **27:** 2257–2267

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E, et al (2012) The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci USA **109:** 11872–11877

Gish LA, Clark SE (2011) The RLK/Pelle family of kinases. Plant J **66:** 117–127

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:** 307–321

Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered **100:** 605–617

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol **30:** 1987–1997

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol **148:** 993–1003

Hok S, Danchin EGJ, Allasia V, Panabières F, Attard A, Keller H (2011) An *Arabidopsis* (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. Plant Cell Environ **34:** 1944–1957

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet **43:** 476–481

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet **41:** 1275–1281

Innan H (2009) Population genetic models of duplicated genes. Genetica **137:** 19–37

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet **11:** 97–108

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature **436:** 793–800

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100

Johnson DA, Thomas MA (2007) The monosaccharide transporter gene family in *Arabidopsis* and rice: a history of duplications, adaptive evolution, and functional divergence. Mol Biol Evol **24:** 2412–2423

Jones DA, Jones JDG (1997) The role of leucine-rich repeat proteins in plant defences. Adv Bot Res **24:** 90–167

Jordan G, Goldman N (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol **29:** 1125–1139

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res **33:** 511–518

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M (2004) Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166:** 935–945

Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet **9:** e1003527

Lehti-Shiu M, Zou C, Shiu SH (2012) Origin, diversity, expansion history, and functional evolution of the plant Receptor-Like Kinase/*Pelle* family. *In* F Tax, B Kemmerling, eds, Receptor-Like Kinases in Plants, **Vol 13**. Springer, Berlin, pp 1–22

Lehti-Shiu MD, Zou C, Hanada K, Shiu SH (2009) Evolutionary history and stress regulation of plant *Receptor-Like Kinase/Pelle* genes. Plant Physiol **150:** 12–26

Leitch AR, Leitch IJ (2012) Ecological and genetic factors linked to contrasting genome dynamics in seed plants. New Phytol **194:** 629–646

Lindner H, Müller LM, Boisson-Dernier A, Grossniklaus U (2012) CrRLK1L receptor-like kinases: not just another brick in the wall. Curr Opin Plant Biol **15:** 659–669

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA **102:** 10557–10562

Markmann K, Giczey G, Parniske M (2008) Functional adaptation of a plant receptor-kinase paved the way for the evolution of intracellular root symbioses with bacteria. PLoS Biol **6:** e68

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature **452:** 991–996

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8:** 122–128

Mühlhausen S, Kollmar M (2013) Whole genome duplication events in plant evolution reconstructed and predicted using myosin motor proteins. BMC Evol Biol **13:** 202

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet **39:** 121–152

Oh DH, Hong H, Lee SY, Yun DJ, Bohnert HJ, Dassanayake M (2014) Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *Schrenkiella parvula*. Plant Physiol **164:** 2123–2138

Oh MH, Wang X, Kota U, Goshe MB, Clouse SD, Huber SC (2009) Tyrosine phosphorylation of the BRI1 receptor kinase emerges as a component of brassinosteroid signaling in *Arabidopsis*. Proc Natl Acad Sci USA **106:** 658–663

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BBH, Jones JDG (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell **91:** 821–832

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551–556

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101:** 9903–9908

Penn O, Privman E, Landan G, Graur D, Pupko T (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol **27:** 1759–1767

Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol **54:** 441–454

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, et al (2012) The cassava genome: current progress, future directions. Trop Plant Biol **5:** 88–94

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna

Renny-Byfield S, Wendel JF (2014) Doubling down on genomes: polyploidy and crop plants. Am J Bot **101:** 1711–1725

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLOS Comput Biol **2:** e115

Romiguier J, Figuet E, Galtier N, Douzery EJ, Boussau B, Dutheil JY, Ranwez V (2012) Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS ONE **7:** e33852

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al (2011) Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res **18:** 65–76

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, et al (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res **15:** 227–239

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463:** 178–183

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA **108:** 4069–4074

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Shang H, Li W, Zou C, Yuan Y (2013) Analyses of the NAC transcription factor gene family in *Gossypium raimondii* Ulbr.: chromosomal location, structure, phylogeny, and expression patterns. J Integr Plant Biol **55:** 663–676

Shiu SH, Bleecker AB (2001) Plant receptor-like kinase gene family: diversity, function, and signaling. Sci STKE **2001:** re22

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KFX, Li WH (2004) Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. Plant Cell **16:** 1220–1234

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Depamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. Am J Bot **96:** 336–348

Soltis DE, Burleigh JG (2009) Surviving the K-T mass extinction: new perspectives of polyploidization in angiosperms. Proc Natl Acad Sci USA **106:** 5455–5456

Song X, Guo P, Li C, Liu CM (2010) The cysteine pairs in CLV2 are not necessary for sensing the CLV3 peptide in shoot and root meristems. J Integr Plant Biol **52:** 774–781

Sun W, Cao Y, Jansen Labby K, Bittel P, Boller T, Bent AF (2012) Probing the *Arabidopsis* flagellin receptor: FLS2-FLS2 association and the contributions of specific domains to signaling function. Plant Cell **24:** 1096–1113

Tan S, Wang D, Ding J, Tian D, Zhang X, Yang S (2011) Adaptive evolution of *Xa21* homologs in Gramineae. Genetica **139:** 1465–1475

Tang H, Bowers JE, Wang X, Paterson AH (2010a) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA **107:** 472–477

Tang P, Zhang Y, Sun X, Tian D, Yang S, Ding J (2010b) Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species. Plant Sci **179:** 399–406

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485:** 635–641

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, et al (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol **30:** 83–89

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al (2010) The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet **42:** 833–839

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet **45:** 487–494

Vetter MM, Kronholm I, He F, Häweker H, Reymond M, Bergelson J, Robatzek S, de Meaux J (2012) Flagellin perception varies quantitatively in *Arabidopsis thaliana* and its relatives. Mol Biol Evol **29:** 1655–1667

Wang GL, Ruan DL, Song WY, Sideris S, Chen L, Pi LY, Zhang S, Zhang Z, Fauquet C, Gaut BS, et al (1998) *Xa21D* encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **10:** 765–779

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet **44:** 1098–1103

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet **43:** 1035–1039

Wu HJ, Zhang Z, Wang JY, Oh DH, Dassanayake M, Liu B, Huang Q, Sun HX, Xia R, Wu Y, et al (2012) Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc Natl Acad Sci USA **109:** 12219–12224

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al (2011) Genome sequence and analysis of the tuber crop potato. Nature **475:** 189–195

Xue Z, Duan L, Liu D, Guo J, Ge S, Dicks J, ÓMáille P, Osbourn A, Qi X (2012) Divergent evolution of oxidosqualene cyclases in plants. New Phytol **193:** 1022–1038

Yang T, Chaudhuri S, Yang L, Du L, Poovaiah BW (2010) A calcium/calmodulin-regulated member of the receptor-like kinase family confers cold tolerance in plants. J Biol Chem **285:** 7119–7126

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:** 1586–1591

Yang Z, Wang Y, Zhou Y, Gao Q, Zhang E, Zhu L, Hu Y, Xu C (2013a) Evolution of land plant genes encoding L-Ala-D/L-Glu epimerases (AEEs) via horizontal gene transfer and positive selection. BMC Plant Biol **13:** 34

Yang ZL, Liu HJ, Wang XR, Zeng QY (2013b) Molecular evolution and expression divergence of the *Populus* polygalacturonase supergene family shed light on the evolution of increasingly complex organs in plants. New Phytol **197:** 1353–1365

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature **480:** 520–524

Yu J, Hu S, Wang J, Wong GKS, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, et al (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nat Biotechnol **30:** 549–554

Zhang XS, Choi JH, Heinz J, Chetty CS (2006) Domain-specific positive selection contributes to the evolution of *Arabidopsis* leucine-rich repeat receptor-like kinase (LRR RLK) genes. J Mol Evol **63:** 612–621

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol **151:** 3–15

BMC Plant Biology

RESEARCH ARTICLE Open Access

Impact of recurrent gene duplication on adaptation of plant genomes

Iris Fischer[1,2*], Jacques Dainat[3,6], Vincent Ranwez[3], Sylvain Glémin[4], Jean-François Dufayard[5] and Nathalie Chantret[1*]

Abstract

Background: Recurrent gene duplication and retention played an important role in angiosperm genome evolution. It has been hypothesized that these processes contribute significantly to plant adaptation but so far this hypothesis has not been tested at the genome scale.

Results: We studied available sequenced angiosperm genomes to assess the frequency of positive selection footprints in lineage specific expanded (LSE) gene families compared to single-copy genes using a d_N/d_S-based test in a phylogenetic framework. We found 5.38% of alignments in LSE genes with codons under positive selection. In contrast, we found no evidence for codons under positive selection in the single-copy reference set. An analysis at the branch level shows that purifying selection acted more strongly on single-copy genes than on LSE gene clusters. Moreover we detect significantly more branches indicating evolution under positive selection and/or relaxed constraint in LSE genes than in single-copy genes.

Conclusions: In this – to our knowledge –first genome-scale study we provide strong empirical support for the hypothesis that LSE genes fuel adaptation in angiosperms. Our conservative approach for detecting selection footprints as well as our results can be of interest for further studies on (plant) gene family evolution.

Keywords: Lineage specific expansion (LSE), Gene duplication, Gene retention, Ultraparalogs (UP), Superorthologs (SO), Comparative genomics, Positive selection, Adaptation

Background

Duplicated genes have been suggested to be the raw material for the evolution of new functions and important players in adaptive evolution [1]. Genomes are constantly subject to rearrangements, by both whole genome duplication (WGD) and small-scale genome duplication (SSD), where tandemly duplicated genes (TDG) are a common case of SSD which generate clusters of physically linked genes. The genomes of angiosperms (flowering plants) are of particular interest to study the impact of gene duplication. Compared to mammals and even to most other plant genomes, angiosperms undergo WGDs, recombination, and retrotransposition more frequently; as a consequence, they also display a larger range of genome sizes and chromosome numbers [2,3]. Most angiosperm genomes sequenced so far show evidence for at least one (but usually more) WGD event during their evolution (see *e.g.* [4-7]). The importance of TDGs has also been shown in *Oryza sativa* (rice) and *Arabidopsis thaliana* where TDGs comprise 15-20% of all coding genes [8-10]. Using genomic and expression data in plants, Hanada *et al.* [11] showed that TDGs tend to be involved in response to environmental stimuli and are enriched in genes up-regulated under biotic stress. This suggests that TDGs play an important role in adaptation of plants to changing environments [11-13]. Taken together, these findings demonstrate the dynamic nature of angiosperm genomes and raise the question of the impact of gene duplications on plant adaptation.

Gene duplication creates an unstable state of functional redundancy, which in most cases will disappear by loss of one copy through accumulation of degenerative mutations, recombination and/or genetic drift. But sometimes both copies are long-term preserved due to functional changes reducing their redundancy and making the loss of one copy disadvantageous [14]. Although the respective roles of adaptive versus non-adaptive processes in the

* Correspondence: irisfischer402@gmail.com; nathalie.chantret@supagro.inra.fr
[1]INRA, UMR 1334 AGAP, 2 Place Pierre Viala, 34060 Montpellier, France
[2]IRD, UMR 232 DIADE, 911 Avenue Agropolis, 34394 Montpellier, France
Full list of author information is available at the end of the article

maintenance of gene duplicates have been much debated (for general reviews see [15-18]), gene duplication should increase the occurrence of adaptation for several reasons. First, it can allow the fixation of beneficial mutations on one copy, leading to neofunctionalization, while the other copy ensures the ancestral function [16,19]. Second, it can free the genome from an "adaptive conflict" if the different functions of an ancestral (single) gene cannot be improved independently [20-22]. Third, even when adaptation is not involved in the initial conservation of duplicates, the presence of two (or more) copies is expected to increase the adaptation rate under certain conditions. Duplication increases the number of gene copies, hence the rate of appearance of beneficial mutations. Otto & Whitton [23] showed that if beneficial mutations are dominant or partly dominant, the rate of adaptation should increase with copy number (or ploidy level). If concerted evolution among gene copies is taken into account, Mano & Innan [24] showed that gene conversion (*i.e.* exchange of genetic material between duplicates in a copy and paste manner) increases the effective population size of gene families proportionally to the number of gene members, thus increasing the efficacy of weak selection. Their model predicts that the rate of adaptive substitutions increases with the number of gene copies. Overall we thus expect higher rates of adaptive evolution in multigene families than in single-copy genes.

As a result of the complex histories of duplicated genes, the retention rate (*i.e.* the proportion of duplicated genes that are maintained in genomes) varies according to several factors including time since the duplication event, protein function, or duplication mode [10]. These variations in retention rates have direct consequences on gene family organization and evolution. Reconciliation methods exploit the observed discrepancies between gene family trees and species trees to infer gene duplication, gene transfer, and gene loss (see [25] for an overview). Among other things, reconciliation methods can be used to estimate duplication or transfer rates and to predict sequence orthology (=sequences related by speciation) [26,27]. Using this method, the extreme heterogeneity of duplication/retention rates among taxa and gene families and/or subfamilies was demonstrated (*e.g.* [28-32]). In particular, reconciliation allows for identification of cases in which recurrent events of duplications (followed by retention) are specific of some lineages and create clades of paralogs (*i.e.* sequences related by duplication) in phylogenetic trees (Figure 1a). Note that since only retained duplications are observable, it is hard to estimate duplication and retention rates independently; hence our use of the "duplication/retention rate" terminology.

Lineage specific duplications/retentions are of particular interest because the recurrence of such events in the same lineage and in a short period of evolutionary time raises the question of their adaptive role to an even greater extent. To test the hypothesis that lineage specific expansion (LSE) of gene families enhances adaptation we compared positive (Darwinian) selection footprints in lineages containing recent and specific duplicated genes to reference lineages containing only single-copy genes. One way to detect positive selection is by analyzing nucleotide substitution patterns at the codon level in a phylogenetic framework.

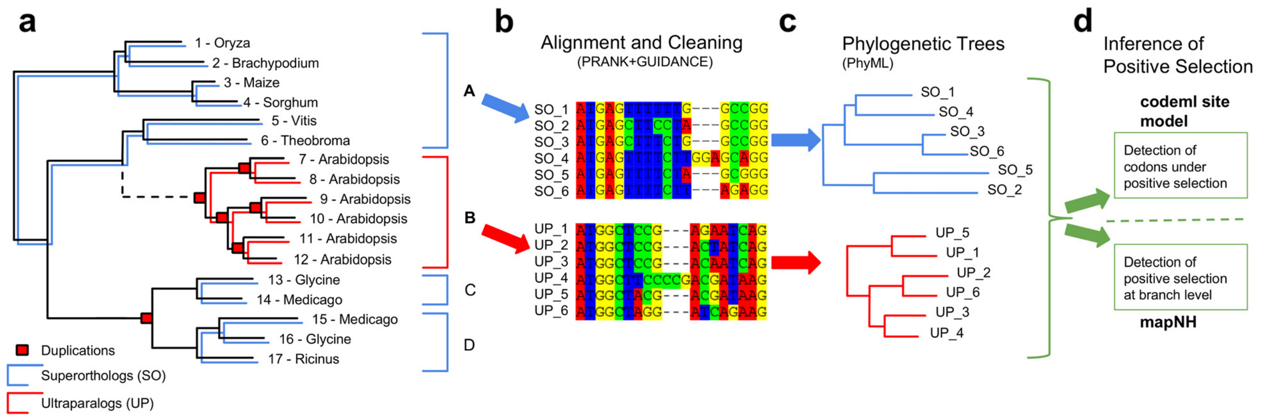


Figure 1 Workflow overview. (a) Example of a protein tree as it can be found on the GreenPhylDB. The Arabidopsis sequences 7 – 12 (cluster B) are only related by duplication (nodes with red boxes) and are therefore ultraparalogs (=UP; red lines). Sequences only related by speciation are superorthologs (=SO; blue lines). Those are sequences 1 – 6 (cluster A), 13 and 14 (cluster C), 15 – 17 (cluster D). For example, sequences 13 and 15 are paralogs (as they are related by duplication) but not ultraparalogs (as a speciation event occurred after duplication). We used only clusters containing at least six sequences (clusters A and B, bold) for further analysis. The dashed lines indicate that SO and UP clusters can come from the same GreenPhyl tree (SO1 and UP1 datasets) or from separate trees (SO2 and UP2 datasets). **(b)** Corresponding CDS sequences were downloaded for each cluster. The clusters were aligned using PRANK [71] and cleaned with GUIDANCE [42]. **(c)** For all alignments, phylogenetic trees were created using PhyML [76]. **(d)** Positive selection was inferred on codons using PAML's codeml [74] and on braches using mapNH [78,79] in all alignments.

Nucleotide substitutions can either be nonsynonymous (*i.e.* protein changing, thereby potentially impacting the fitness) or synonymous (*i.e.* not protein changing, thereby theoretically without consequences for the fitness). The nonsynonymous/synonymous substitution rate ratio, denoted as d_N/d_S or ω, can be used to infer the direction and strength of natural selection. If no selection is acting, ω should equal 1. An ω value smaller than 1 indicates an under-representation of nonsynonymous substitutions, which can be interpreted as the preferential elimination of deleterious mutations by purifying selection. The closer ω is to zero, the stronger purifying selection is acting. On the other hand, if ω is larger than 1 it indicates an over-representation of nonsynonymous substitutions, which can be interpreted as positive selection on new variants. Using such an approach, positive selection has been detected for MADS-box transcription factors [33], monosaccharide transporters [34], genes involved in a triterpene pathway [35], an anthocyanin pathway enzyme encoding gene [36], and epimerase genes [37] to mention only a few examples in plants. So far, this approach has mostly been applied to single candidate gene families. Thanks to the availability of numerous completely sequenced plant genomes, it can now be used at the genome level for several angiosperm species.

The dynamic nature of angiosperm genomes makes them an ideal system to study the link between gene duplication/retention rate heterogeneity and adaptation. Assuming that adaptation is acting when positive selection footprints are detected, we want to test if positive selection can be observed more frequently in LSE genes compared to single-copy genes. We applied a d_N/d_S-based test to detect positive selection as it is easy to use on a large scale, it is one of the most stringent tests [38-40], and it has been applied successfully in many similar cases (for examples, see above). Using this approach, we found 5.38% of codons under positive selection in LSE gene families but none in single-copy ones. In addition, the average ω over branches of LSE gene trees is almost twice as high as that observed in single-copy gene trees. We also found a much higher proportion of branches under positive selection and/or relaxed constraint among LSE gene trees than among single-copy gene trees. Taken together, these results strongly support the prediction that (at least in angiosperm genomes) LSE gene evolution plays an important role in adaptation whereas very few single-copy genes seem to be involved.

Results

Dataset description

We investigated whole genomes of five monocots (*Musa acuminata*, *O. sativa*, *Brachypodium distachyon*, *Zea mays*, *Sorghum bicolor*) and five dicots (*Vitis vinifera*, *A. thaliana*, *Populus trichocarpa*, *Glycine max*, *Medicago truncatula*). From the GreenPhyl database [41] we extracted ultraparalog clusters (UP – sequences only related by duplication) which represent our LSE gene set. As a single-copy gene reference, we chose a superortholog gene set (SO – sequences only related by speciation). To address the question of whether or not positive selection is more frequent during LSE events, we compared the results obtained on UPs with those obtained on SO gene sets. The SO gene set was then divided in two subsets. The first one, SO1, contains SO genes extracted from GreenPhyl protein trees in which at least one UP cluster was also identified. This means that all the trees from which an SO1 was extracted contain at least one UP cluster. The second SO set (SO2) is the complement of SO1, *i.e.* it is composed of SO genes extracted from GreenPhyl trees in which no UP clusters were found. Likewise, the UP1 dataset represents UP clusters extracted from GreenPhyl trees also containing SO clusters and the UP2 dataset represents UP clusters from GreenPhyl trees from which no SO clusters were extracted. We subdivided the dataset as we expected a "family effect". This effect may be caused by an accelerated evolutionary rate in some families which are more prone to gene duplication and/or retention than others, *e.g.* due to their function or base composition. If one GreenPhyl tree contained more than one SO or UP cluster, we kept only one cluster randomly (see Methods for details). A detailed overview of the workflow can be found in Figure 1.

Our final dataset for codeml analysis comprised 160 UP1, 1,512 UP2, 167 SO1, and 1,203 SO2 clusters (Table 1). The mapNH analysis was performed on 154 UP1, 1,435 UP2, 167 SO1, and 1,203 SO2 clusters (Table 1) and 1,257 UP1, 14,326 UP2, 1,807 SO1, and 13,374 SO2 branches (Table 1). The median length of the UP1 alignments is 1,272 bp (base pairs), 1,220 bp for the UP2, 1,230 bp for SO1, and 987 bp for SO2 alignments (Table 1, Figure 2). The UP alignments are significantly longer than the SO alignments (Mann–Whitney test: $p < 0.001$). This can be partially explained by the fact that GUIDANCE introduces gaps instead of aligning ambiguous sites [42]. Therefore, UP genes – which are frequently under less selective constraint – may produce longer alignments due to the introduction of gaps. The median number of sequences in an alignment (*i.e.* median cluster size) is 7 for UP and SO alignments (Table 1, Figure 2). We found that the cluster sizes for the SO datasets are significantly smaller than for the UP datasets (Mann–Whitney test: $p < 0.001$) which was expected because the number of sequences a superortholog cluster can contain is at most ten (=number of species used in this study) whereas it is not bounded for UP clusters.

As this divergence time between one species and its closest relative increases, one might expect that the

Table 1 General dataset description

	UP1	UP2	UPps	SO1	SO2
Clusters for final codeml site model analysis	160	1,512	90	167	1,203
Clusters for final mapNH analysis	154	1,435	90	167	1,203
Total number of branches	1,881	22,475	1,730	1,817	13,537
Number of analysed branches by mapNH	1,257	14,326	1,298	1,807	13,374
Median cluster size (1st Qu; 3rd Qu)	7 (6; 8)	7 (6; 10)	8 (6; 13)	7 (6; 8)	7 (6; 8)
Median alignment length (1st Qu; 3rd Qu) [bp]	1,272 (792; 1,858)	1,220 (753; 1,851)	1,314 (864; 1,942)	1,230 (900; 1,737)	987 (651; 1,470)
Total number of branches	1,881	22,475	1,730	1,817	13,537
Total number of sites	42,706	355,486	21,864	59,191	340,556

Qu quantile; *bp* base pairs.

number of detected UPs could also increase when compared to a distantly related species than to a closely related one. Therefore, we tested if the divergence time and the number of identified UP clusters correlated. Note that we always used the divergence time relative to the most closely related species in the GreenPhyl database, no matter if we analysed this species later (divergence times can be found in Additional file 1: Figure S1). Regression analysis shows that there is no significant positive correlation between the divergence time and the number of detected clusters: Spearman non-parametric correlation coefficient (ρ) = −0.171, $p = 0.626$ (Figure 3). The correlation remains not significant after removing *M. trunculata* ($\rho = -0.227$, $p = 0.557$). The most likely explanation for this lack of correlation is the equilibrium between gene duplication and loss over time. The birth/death rate has been shown to be relatively constant over time and therefore the frequency of gene copies in a genome declines exponentially with age [14].

Positive selection at the codon level

The average number of UP clusters used in the final analysis is around 150 clusters per species, with *Brachypodium distachyon* showing a very low (63) and *Medicago truncatula* showing a very high (400) number of clusters (Table 2). On average, 12.86% and 5.38% of UP clusters show evidence for positive selection before and after manual curation, respectively (Table 2). This discrepancy shows how important manual curation for alignment errors is as we discovered around 50% of alignments with a possible false positive signal. As we were very strict during the manual curation process, the clusters remaining can be considered as true positives but we might have removed some other true positives. There is no significant

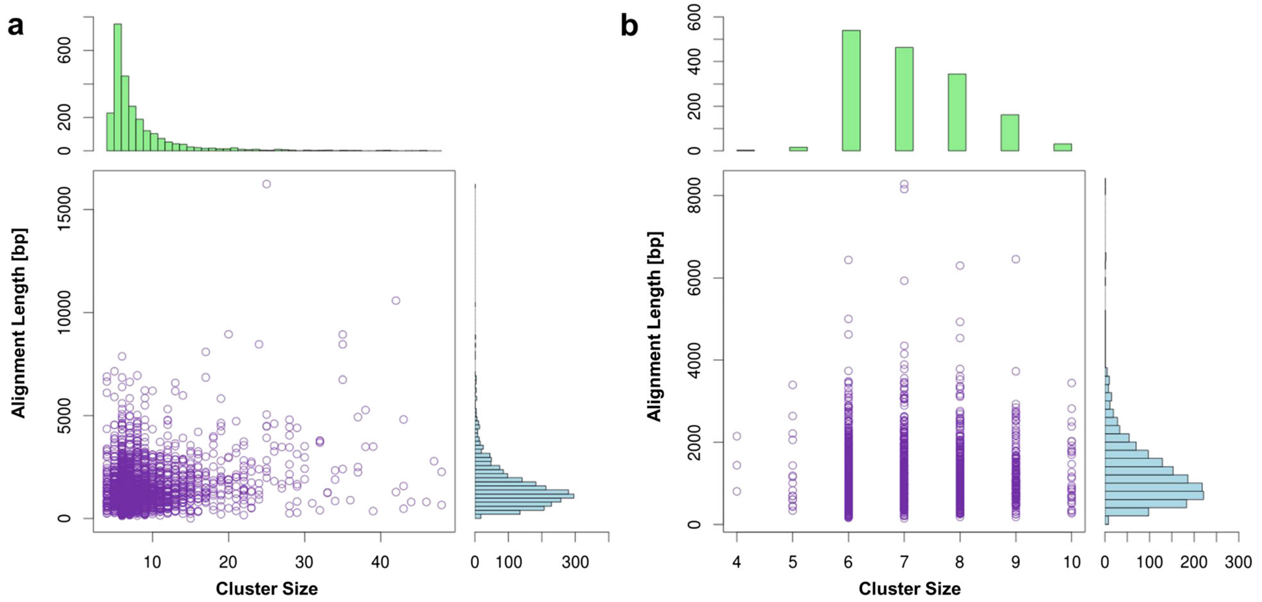


Figure 2 Alignment length against cluster size. Each dot in the scatter plot represents **(a)** an ultraparalog or **(b)** a superortholog alignment. The histogram above the scatter plot represents the count of alignments for each cluster size; the histogram right to the scatter plot represents the counts of alignments for each alignment length. bp: base pairs.

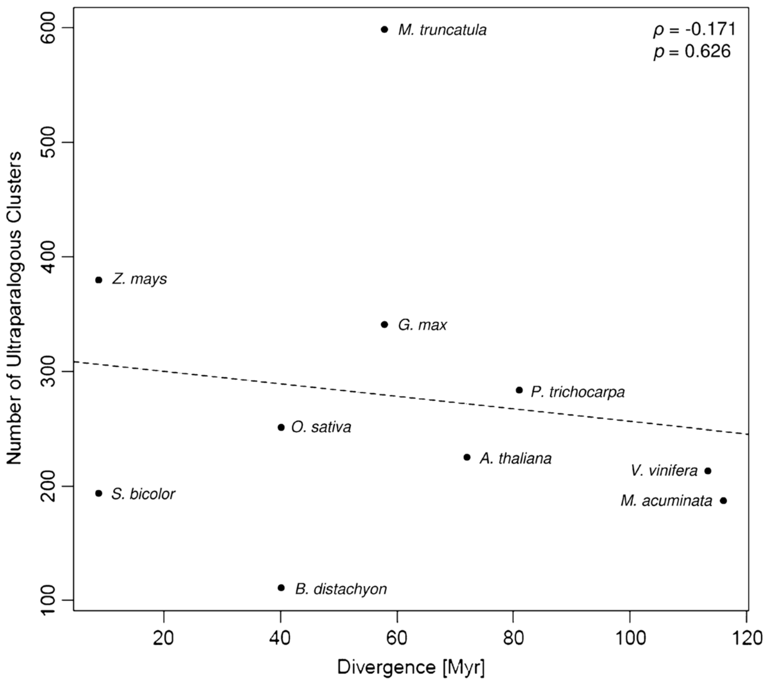


Figure 3 Number of detected UP clusters for every species against divergence time. No significant correlation was observed (ρ: −0.171, *p* = 0.626).

difference between the number of UP1 and UP2 clusters under selection although we detected less – sometimes zero – clusters with codons under selection in the UP1 dataset, most likely because of a small sample size in this dataset (160 clusters vs. 1,512 UP2 clusters). Interestingly, no SO1 or SO2 cluster seems to have evolved under positive selection (Table 2). We also defined a new subcategory of clusters denoted UPps that contains the 90 UP clusters for which positive selected sites were detected and manually validated (Table 1). The UPps clusters have a longer median length (1,314 bp) and larger median cluster size (8) than the other UP and SO clusters (Table 1).

Table 2 Clusters containing codons under positive selection before and after manual curation

	Clusters used in final analysis		Clusters under selection before manual curation (%)		Clusters under selection after manual curation (%)	
Species	**UP1**	**UP2**	**UP1**	**UP2**	**UP1**	**UP2**
M. acuminata	36	107	1 (2.78)	6 (5.61)	0 (0.00)	4 (3.74)
O. sativa	7	145	1 (14.29)	29 (20.00)	1 (14.29)	11 (7.59)
B. distachyon	4	59	0 (0.00)	14 (23.73)	0 (0.00)	2 (3.39)
Z. mays	24	226	4 (16.67)	32 (14.16)	0 (0.00)	9 (3.98)
S. bicolor	4	93	0 (0.00)	9 (9.68)	0 (0.00)	4 (4.30)
V. vinifera	9	114	1 (11.11)	10 (8.77)	0 (0.00)	3 (2.63)
A. thaliana	13	138	0 (0.00)	25 (18.12)	0 (0.00)	14 (10.14)
P. trichocarpa	16	132	3 (18.75)	18 (13.64)	1 (6.25)	12 (9.09)
G. max	17	128	3 (17.65)	5 (3.91)	3 (17.65)	1 (0.78)
M. truncatula	30	370	5 (16.67)	49 (13.24)	4 (13.33)	21 (5.68)
Sum/average	*160*	*1,512*	*18 (11.25)*	*197 (13.03)*	*9 (5.63)*	*81 (5.36)*
UPall	*1,672*		*215 (12.86)*		*90 (5.38)*	
SO1	167		1 (0.60)		0 (0.00)	
SO2	1,203		3 (0.25)		0 (0.00)	

ω at the branch level

The analysis of selective pressures at the branch level was performed using mapNH on the same dataset as the codon analysis. If ω at a branch is larger than 1.2 we consider this a strong indicator of positive selection (simply defining $\omega > 1$ as an indicator of positive selection might lead to false positives as in a neutral scenario ω rather fluctuates around 1 than being exactly 1). The mean ω of the branches is significantly ($p < 0.001$) higher in UP2 (0.62) than in SO2 (0.29) and the distribution shows a larger variance for UP2 than for SO2 (Figure 4, Table 3). As compared to SO2, in UP2 we observe: (i) a higher proportion of branches with $\omega > 1.2$ (8.78%, compared to 0.22% for SO2), (ii) higher ω values for branches with $\omega > 1.2$ (1.80, compared to 1.64 for SO2), and (iii) higher ω values for branches with $\omega < 1$ (0.49 compared to 0.29 for SO2; Table 3). This indicates a relaxation of purifying selection for UP2 in contrast to SO2 but also a higher frequency of branches harboring an accelerated evolution rate. Similar results are observed on the UP and SO clusters extracted from the same trees (*i.e.* UP1 and SO1). Mean ω is significantly ($p < 0.001$) higher for UP1 (0.51) than for SO1 (0.28; Table 3). Interestingly, the mean ω for UP1 and UP2 differ significantly ($p < 0.001$; Table 3, Figure 4), indicating the family effect mentioned before. For the UPps clusters, the mean ω (0.84), the proportion of branches with $\omega > 1.2$ (15.79%), and the mean ω of branches with $\omega > 1.2$ (1.95) are higher compared to the UP1 and UP2 clusters (Table 3, Figure 4).

Effect of cluster size and length

The UP clusters are longer and contain more sequences than the SO clusters (see above). This could lead to an underestimation of codons under selection in SO clusters as codeml has more power to detect footprints of positive selection in longer/larger alignments [39]. A general linear model analysis showed that differences in alignment length cannot explain the detected differences between UP and SO clusters if cluster size (=number of sequences in alignment) is ≤ 10 (data not shown). Cluster size, however, had an effect. In order to test the reliability of our results relative to the number of sequences, we performed Fisher's exact tests to see if we could find either significantly more clusters, codons, and/or branches under selection for UP than for SO cluster in each cluster size category (up to 10 as this is the maximum for SO clusters). We find significantly more clusters under positive selection for UP clusters for the size categories 6 and 7 (Table 4). For the other size categories we lack power to detect significant differences (Table 4). We also detect significantly more codons showing footprints of selection in UP clusters for the size categories 6–9 (Table 4). In addition, branches with $\omega > 1.2$ are significantly more frequent in UP clusters for size categories 5–10 (Table 4). To summarize, UP clusters still show more signatures of positive selection more frequently after controlling for cluster size effect.

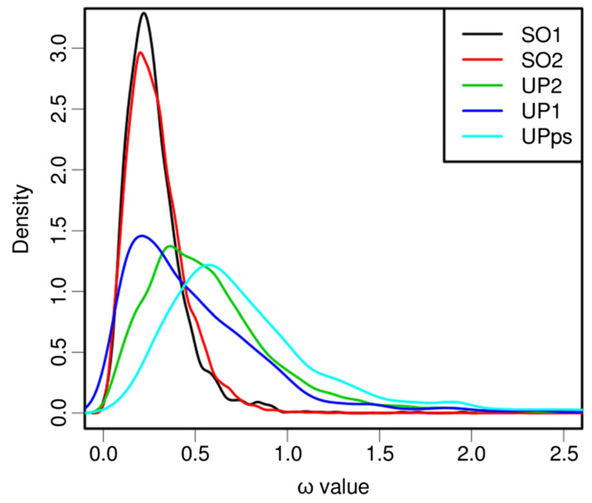


Figure 4 Distribution of ω of branches in different subsets. Distribution of ω of branches in SO1 (black), SO2 (red), UP1 (green), UP2 (dark blue), and UPps (light blue) clusters.

Effect of evolutionary time and polymorphism

To see if our results are biased by divergence discrepancies between UP and SO, we sorted the ω value of each branch by their synonymous substitution rate (d_S). To rule out the effect of polymorphism, we excluded ("young") external branches from the dataset and compared the remaining ("old") internal branches (UPint) to the SO dataset. We found a significant difference between the ω of SO and UPint in d_S intervals ranging from 0.01 to 0.21 (Figure 5a+b). There is no significant difference in the first d_S interval (Figure 5a), most likely because of residual polymorphism and/or a low mutation rate in SO and UP clusters. This interval harbors, however, more than 50% of the dataset. These results indicate that – except for very low d_S values – the difference between SO and UP cluster cannot be explained solely by divergence discrepancies or residual polymorphism. Above d_S values of 0.21 the Mann–Whitney test is inconclusive (Figure 5b) due to lack of power.

Annotation of clusters under selection

The GreenPhylDB provides details on predicted molecular function, biological process, cellular component, and family and domain annotation for each cluster. We extracted those details for clusters found to have evolved under positive selection using codeml's site model. Additional file 2 provides the details of the annotations for all the clusters with codons under selection before excluding clusters which derive from the same GreenPhyl tree (see Methods).

Table 3 Results of the branch analysis with mapNH

	UP1	UP2	UPps	SO1	SO2
Number of analysed branches	1,257	14,326	1,298	1,807	13,374
Branches with $\omega < 1$ (%)[a]	1,144 (91.01)	12,515 (87.36)	993 (76.50)	1,799 (99.56)	13,329 (99.66)
Mean ω for branches with $\omega < 1$	0.41	0.49	0.59	0.28	0.29
Branches with $\omega > 1$ (%)[a]	113 (8.99)	1,811 (12.64)	305 (23.50)	8 (0.44)	45 (0.34)
Mean ω for branches with $\omega > 1$	1.55	1.52	1.67	1.37	1.44
Branches with $\omega > 1.2$ (%)[a]	73 (5.81)	1,099 (8.78)	205 (15.79)	4 (0.22)	23 (0.17)
Mean ω for branches with $\omega > 1.2$	1.81	1.80	1.95	1.64	1.79
Mean $\omega \pm$ SE	0.51 ± 0.44	0.62 ± 0.47	0.84 ± 0.65	0.28 ± 0.17	0.29 ± 0.17

[a] of analysed branches.

Annotation is an ongoing process on the GreenPhylDB; therefore most of the clusters are not annotated – especially in monocots. There seems to be no trend in tree size or species specificity as clusters shown to have codons under selection can both be found in large trees containing sequences from various plant species and from small species specific trees (Additional file 2). As annotation is ongoing and remains under constant modification, a comprehensive analysis of the potential function of the clusters with codons under selection would not lead to reliable results. However, some trends can be observed: (i) the most abundant molecular function is "protein binding" (21.57% of all annotated molecular functions in the dataset) followed by "transferase activity" (9.80%). This is especially true in the Level 2 dataset (*i.e.* clusters derived from large GreenPhyl trees) whereas potential molecular functions seem to be more diverse in the Level 1 dataset (Additional file 2). (ii) The most common predicted biological functions are "metabolic process" (23.53% of all annotated biological processes in the dataset) and "oxidation-reduction process" (20.59%). "Defense" (14.71%) is also dominant, but only in the Level 2 dataset (Additional file 2). (iii) If domains are annotated to the clusters with codons under selection, F-box (22.54% of all annotated domains in the dataset), Leucine rich repeats (LRR; 11.27%), and NB-ARCs (8.45%) are predominant. Again, this trend is mostly visible in the Level 2 dataset whereas potential domains are more diverse in the Level 1 dataset (Additional file 2).

Discussion

The important role of duplicated genes in plant adaptation has been argued theoretically (reviewed by [43]). To assess whether lineage specific expanded (LSE) genes show more evidence for positive selection than single-copy genes we analyzed LSE gene families from ten angiosperm genomes using a d_N/d_S-based test. We found positive selection footprints moderately frequently at the codon level in LSE genes (5.38% in average among the different species) but did not find any positive selection footprints on single-copy genes after manual curation. The number of codons under positive selection is also found higher in LSE than in single copy genes for different cluster size categories and thus cannot be explained solely by a difference of power to detect positive selection between the two datasets. Positive selection is also detected in LSE genes at the branch level and we found a significantly higher

Table 4 Results of Fisher's exact test

	Number of clusters			Number of codons			Number of branches		
Cluster size	UP under/not under positive selection	SO under/not under positive selection	p-value Fisher's exact test[a]	UP under/not under positive selection	SO under/not under positive selection	p-value Fisher's exact test[a]	UP under/not under positive selection	SO under/not under positive selection	p-value Fisher's exact test[a]
4	1/48	0/3	1	2/22,821	0/1,467	1	9/210	0/15	1
5	4/102	0/12	1	16/51,017	0/4,187	0.62	51/483	0/84	8.78E-04[*]
6	24/474	0/487	9.27E-08[***]	66/210,761	0/191,533	2.20E-16[***]	184/3,456	6/4,329	2.20E-16[***]
7	15/280	0/405	1.90E-06[***]	43/127,403	0/163,947	3.59E-16[***]	136/2,299	8/4,378	2.20E-16[***]
8	4/178	0/293	0.02	24/81,803	0/110,494	1.24E-09[***]	117/1,430	5/3,744	2.20E-16[***]
9	7/108	0/144	3.07E-03	19/49,429	0/57,346	4.42E-07[***]	69/1,110	7/2,085	2.20E-16[***]
10	4/73	0/26	0.57	14/36,324	0/10,298	0.05	76/714	1/425	6.03E-14[***]

The table contains the results of Fisher's exact test for number of clusters, codons, and branches under positive selection vs. not under positive selection in UP and SO clusters for different cluster size categories.

[a] Bonferroni corrected for multiple testing.

[*] p < 2.38E-03, [***] p < 4.76E-05.

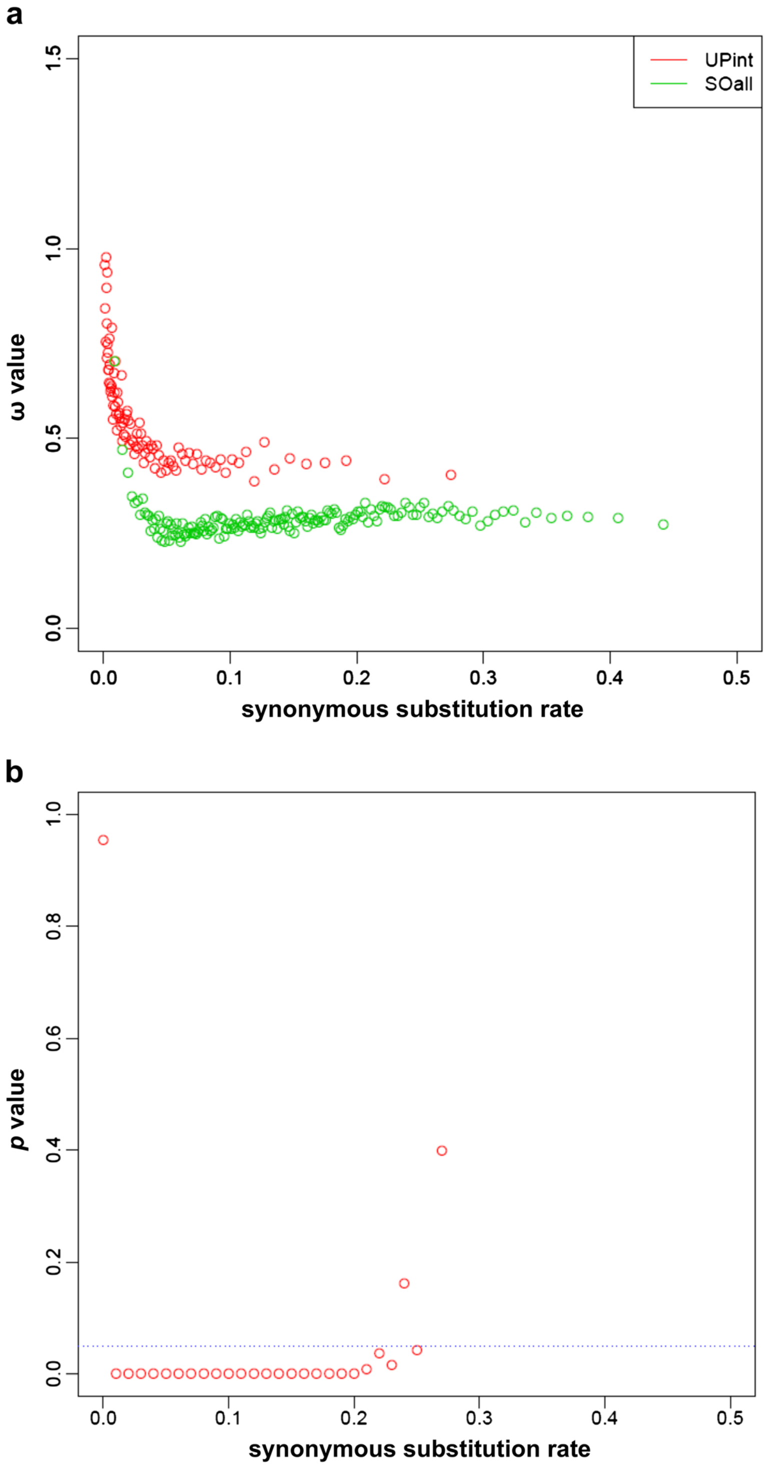


Figure 5 ω of branches according to the ratio of synonymous mutations. (a) ω of internal branches of UP clusters (red) and all branches of SO clusters (green) is plotted against the rate of synonymous mutations of sequences. As the point density is too high, each point represents the mean of 100 values. **(b)** The *p*-value of the Mann–Whitney test according to the synonymous substitution rate. This statistical test is performed using all the ω data and on intervals of 0.01 and contains at least 25 values. The dotted blue line is the significance level fixed at 0.05.

proportion of branches under positive selection among LSE gene trees than among single-copy ones. Inferring d_N/d_S at the branch level is complementary to analyzing d_N/d_S at the codon level. Using site models, d_N/d_S-based tests have the greatest power to detect footprints of selection in genes involved in co-evolutionary processes as a limited subset of their codons is repeatedly subject to positive selection (reviewed by [44]). At the branch level, the evolutionary rate is averaged over the complete amino acid sequence, making it difficult to detect a signal when only few sites are targets of positive selection. However, an elevated evolutionary rate can be detected even if it affects only certain lineages. When d_N/d_S was computed on all the branches of the same dataset as for site analyses, we detected a stronger effect of positive selection on LSE genes compared to single-copy genes. Therefore, we argue that LSE genes are a much more important substrate for positive selection to act on than single-copy genes. This is – to our knowledge – the first genome-scale study to empirically demonstrate that LSE genes fuel adaptation in angiosperms.

Among the vast literature dealing with population genetic models of duplicated gene evolution, a crucial point is whether natural selection plays a role in it [18]. Positive selection is expected to act either on the fixation process of the duplication itself or at new mutations occurring after fixation of the copy in the species (or at both levels successively). We found a significantly larger portion of LSE genes under positive selection compared to single copy ones. Hence, the differentiation between copies for LSE genes is driven by changes in proteins, with all the functional consequences this may imply. This result corresponds to predictions made by several models, *e.g.* the "adaptation" model [16,19] or the "adaptive conflict" model [20-22]. In these scenarios, the duplication itself is not subject to positive selection, and may be fixed by genetic drift. However, our results may be coherent with a third scenario of segregation avoidance [45] where several alleles are pre-existing at the ancestral unique locus and their retention is advantageous [46,47]. Thus, duplications may favor the retention of those alleles if each of them gets fixed at one of the different locus resulting from the duplication process. In this scenario, positive selection does occur on the fixation process itself and the non-synonymous mutation observed would have appeared before the duplication process. However, it is not possible to tell which of these scenarios is more likely in our data, all the more that those scenarios can be combined in more complex ones. For instance, a first duplication may occur allowing a unique gene to escape an adaptive conflict and subsequent duplications may occur; generating additional copies following – this time – an adaptive scenario.

Recent progress in angiosperm whole genome sequencing gave numerous arguments in favor of the positive role of polyploidy in the exceptional radiation and diversification of angiosperms [48-50]. These hypotheses rely on the evolutionary potential caused by genomic shocks such as polyploidy. Our study shows that genomic events leading to gene duplications at a smaller scale – especially when recurring at a high frequency as it has been described in angiosperm genomes [8-10] – appear also fundamental in the adaptive dynamic of angiosperms. Recurrent gene duplication/retention offer a mechanism complementary to WGD as it may take place all along the evolutionary time and can affect a specific subset of gene families. Such families might be targeted according to their implication in biological processes or molecular functions related to the ongoing natural selective pressure. This could be reflected by the trends we observed in the annotations of the genes containing codons under selection: many are involved in defense and protein binding is the most common molecular function.

The most abundant domains we found in LSE clusters showing signatures of positive selection are F-box and LRR domains. F-box proteins (FBP) are one of the largest and fastest evolving gene families in land plants [51]. When analyzing FBP subfamilies in seven land plant species, it was found that 64-67% of duplications are species-specific – mostly in angiosperms [52,53]. Expression analysis of LSE FBPs showed a fast subfunctionalization on the transcriptional level [52,53]. Finally, it was also found that the LSE FBP are less conserved than their single-copy counterparts and signatures of positive selection are predominantly found in the protein-protein interaction domains of the FBPs [52,53]. An equally large gene family comprises of receptor-like kinases (RLK) containing LRRs in their extracellular domain [54]. Two main functions are described for LRR-RLKs: development and defense [55]. LRR-RLKs involved in defense are predominantly found in LSE gene clusters whereas LRR-RLKs involved in development are mostly found in non-expanded groups [55]. It was also discovered that the LRR domains are significantly less conserved than the remaining domains of the LRR-RLK genes [55]. In addition, a study on four plant genomes showed that LRR-RLK genes from LSE gene clusters show significantly more indication of positive selection or relaxed constraint than LRR-RLKs from non-expanded groups [55]. Therefore, it is not surprising that F-box and LRR domains are the most abundant domains we found in the LSE clusters with codons under positive selection. First, proteins containing these domains constitute large gene families and are therefore likely to show up in our LSE dataset – especially when coming from the GreenPhyl Level 2 dataset as it comprises of large trees. Second, several studies showed that these proteins/domains are prone to fast evolution and adaptation [51,55]. The results shown here give valuable insight in the evolution of large gene families and provide the groundwork for more detailed analyses of these candidates.

As automated multi-step genome wide analyses can sometimes introduce biases and misinterpretations, we took the maximum of precautions at each step. First, we chose well-annotated genomes to reduce the bias of misannotations, although we cannot completely rule them out. Annotation errors could lead to an over-estimation of the evolutionary rate in duplicated genes [56]. This left us with ten angiosperm genomes, even though many completely sequenced genomes are now available. Second, as d_N/d_S-based methods are very sensitive to alignment errors [57,58], reliable alignment and cleaning tools are mandatory. We used PRANK and GUIDANCE to align and clean the sequence clusters. Those recent methods have been found to produce the most reliable alignments for downstream analysis using the PAML software [57,58]. Third, we curated the alignments for which we detected positive selection manually. As this is a great deal of work in large datasets many studies fail to do this. However, we argue that this step is crucial to produce reliable results as we found around 50% alignment errors and therefore false positives. The manual validation of all the positively selected sites is a major strength of our study. Fourth, the power for d_N/d_S analysis is related to the number of sequences aligned. In our dataset the difference in sequence number was significant between the LSE and the single-copy dataset. This could explain, at least partially, the detection of a higher number of clusters with sites under positive selection. By analyzing LSE and single-copy gene clusters in each size categories separately we ruled out the effect of cluster size and showed that the number of clusters, codons and branches under positive selection is always higher in LSE genes compared to single-copy genes. Fifth, we wanted control for a potential "family effect" that could result from the fact that some gene families showing accelerated evolutionary rate in general, *e.g.* because of their function or base composition, may also be more prone to gene duplication and/or retention than others. Using subgroups we indeed found an effect: LSE clusters from trees containing also a single-copy gene clusters show a lower d_N/d_S compared to LSE clusters from trees without single-copy gene clusters. This means that the more a gene family is prone to duplication/retention the less probable a single-copy gene cluster will be found. Here, we give an argument in favor of the hypothesis that the initial level of selective constraint partially conditions the frequency of duplication/retention. We detect a family effect in different trees but the d_N/d_S difference between LSE clusters and single-copy gene sets remains significant when controlling for this effect by comparing clusters extracted from the same gene trees.

Finally, when analyzing very recent duplicates it is possible that the differences between copies are still segregating within populations which violates basic assumptions of d_N/d_S-based tests [59]. Our LSE dataset may include genes where differences are still polymorphic which can lead to an overestimation of positive selection [59,60]. As expected, d_N/d_S is elevated – and most likely overestimated – for low d_S values in LSE as well as in single-copy gene clusters. The reason for this effect is either polymorphism segregating in young copies (mostly the case in LSE genes) or a low mutation rate (mostly the case in single-copy genes). However, even after removing external ("young") LSE branches, the difference between single-copy and LSE gene clusters is still significant for d_S values above 0.01. This result shows that polymorphism and/or a low mutation rate alone cannot explain the differences in d_N/d_S between LSE and single-copy genes.

Functional analysis is difficult in recently expanded gene families because functional or gene expression differences are difficult to investigate due to highly similar sequences among copies. Additionally, many of these genes are involved in stress responses [11,12] and therefore specific conditions need to be defined *a priori*. Consequently, molecular evolution studies like ours are a good alternative to identify candidates in which family expansion is followed by an adaptive process to conduct further analyses. Another next step could be to investigate links between our results and the duplication mode. By looking at the location of duplicated genes in the genome the duplication mode can be assessed. Several studies showed that the duplication mode has an impact on genetic novelty and adaptation [61,62]. For example, it was demonstrated that TDGs are more often involved in abiotic stress response than non-TDGs [10,11,63]. However, a d_N/d_S approach is not suitable to provide evidence for positive selection on the duplication process itself which is the assumption under the dosage effect hypothesis [13]. Therefore, we ignore gene conservation as potential outcome and subsequently probably underestimate the role of adaptation in gene duplication/retention.

Conclusions

In this paper we conduct one of the largest studies on the role of recurrent gene duplication on adaptation in angiosperms so far. Indeed, most of the former studies either dealt with candidate families in a broad taxonomical range (*e.g.* [35-37]) or whole genomes for a maximum of four plant species (*e.g.* [11,12]). We searched duplicated genes from ten angiosperm genomes for footprints of positive selection and our results provide candidates for further functional or population genetic studies. In general, we used a very conservative approach to detect positive selection footprints at LSE genes and might therefore miss many true positives. Still, because of the inherent differences between LSE and single-copy datasets, our results must be interpreted with caution. As the number and quality of sequenced genomes is increasing daily, our analysis can be expanded to many more plant species in the future. In addition, current efforts in re-sequencing numerous genomes from different populations could give the opportunity to differentiate between divergence and polymorphism and to consequently provide even better estimates of quantity and quality of positive selection undergone by LSE genes.

Methods

Genomes, proteomes, identification of ultraparalog clusters and superortholog gene sets

As analysis of duplicated genes is very sensitive to gene annotation errors we chose five well annotated monocot and five well annotated dicot genomes (see details on our genome selection criteria in Additional file 1): *Musa acuminata* v1.0 (banana) [5], *Oryza sativa* subsp. *japonica* v6.0 TEfiltered (Asian rice) [9], *Brachypodium distachyon* v1.0 (purple false brome) [64], *Zea mays* v5.6 filtered (maize) [65], *Sorghum bicolor* v1.4 (milo) [66], *Vitis vinifera* v1.0 (common grape vine) [4], *Arabidopsis thaliana* v10.0 (thale cress) [8], *Populus trichocarpa* v2.2 (black cottonwood) [67], *Glycine max* v1.0 (soybean) [68], and *Medicago truncatula* v3.5 (barrel medic) [69]. The phylogeny of those species is provided in Additional file 1. We used the information provided by the GreenPhyl v3 database (http://www.greenphyl.org) which uses a tree reconciliation approach [70] to identify orthologs (genes related by speciation) and paralogs (genes related by duplication) in protein trees. This database contains protein families' composition and phylogenies for a broad variety of green plants whose genomes have been completely sequenced [41]. Based on their sequence similarity, the GreenPhylDB clusters gene families at different levels from the less stringent (large clusters of relatively similar sequences at Level 1) to the most stringent (small clusters of highly similar sequence at Level 4). First, we extracted 3,330 protein clusters from Level 1. As large gene families (>500 sequences) are not further analyzed in GreenPhyl, we extracted 2,238 protein clusters from Level 2 for these gene familie. These are two separate datasets and Level 2 trees are not nested in Level 1 tress (see GreenPhyl homepage for details: http://www.greenphyl.org/).

We extracted ultraparalog clusters (UP – sequences only related by duplication) from the GreenPhylDB trees on which duplication and speciation events were positioned according to the tree reconciliation approach cited previously (Figure 1a). Those clusters represent our LSE gene set. As a single-copy gene reference, we chose a superortholog gene set (SO – sequences only related by speciation). We ignored clusters with less than six sequences. The SO clusters were divided into clusters coming from the same tree as UP clusters (SO1) or from trees exclusively harboring SO clusters (SO2). Likewise, UP clusters were divided in clusters coming from trees containing SO clusters (UP1) or from trees with only UP clusters (UP2). Note that when a GreenPhyl tree harbors several SO and/or UP clusters, all were extracted. We downloaded the corresponding complete CDS of the species of interest (links on GreenPhylDB Documentation section). In case of alternative spliceforms, the longest one is kept in the GreenPhylDB pipeline; it is thus the one we downloaded. Most GreenPhyl trees are too large and/or too divergent to create reliable nucleotide alignments and perform d_N/d_S-based tests on the whole tree alignment. This is especially true for the most interesting cases where trees contain both UP and SO clusters (the UP1/SO1 dataset). We therefore chose to analyze each UP and SO cluster independently.

In GreenPhyl trees harboring several UP and/or SO clusters *i.e.* in gene families in which gene duplication/retention might be more frequent, one might expect selective constraint to be different, in particular more relaxed. Therefore, some gene families might be overrepresented when several clusters from the same tree are analyzed separately. To avoid this, an additional step of selection was added to the initial dataset as we randomly kept only one cluster each time several clusters of UP or several clusters of SO were identified from a same tree and removed all other clusters from our analysis. Here, we present the results for this final sub-dataset. However, we performed our analysis on three additional sub-datasets: (i) the whole dataset without removing clusters from trees harboring more than one cluster, (ii) a dataset which contains clusters from GreenPhyl trees with only one UP and/or one SO cluster, (iii) a dataset where only clusters from trees harboring more than one cluster were kept. The results for these sub-datasets can be found in Additional file 1. However, the trends we observe remain, no matter which sub-dataset is analyzed (Additional file 1).

Alignment and cleaning

We used PRANK_{+F} with codon option [71] for creating the alignments and GUIDANCE [42] with the default sequence quality cut-off and a column cut-off of 0.97 to remove problematic sequences and unreliable sites from the initial alignments (Figure 1b). Those choices were guided by several recent studies which found $\text{PRANK}_{\text{codon}}$ and the $\text{PRANK}_{\text{codon}}$-GUIDANCE combination to produce the most reliable alignments for further inference of positive selection using codeml [57,58]. Filtering removed all sequences from 33 UP clusters, it kept three or less sequences for 91 UP and two SO clusters; all those clusters were thus ignored in further analyses as a minimum of four sequences was required. For some species (namely *Z. mays*, *S. bicolor*, *G. max*, and *M. truncatula*), the retrieved CDS seemed to contain un-translated regions (UTRs) as for 126 UP and four SO clusters one or more sequences contained stop codons or incomplete codons (*i.e.* length not divisible by three). Those clusters were also removed from the analysis. Additionally, for 18 UP clusters codeml failed to run (probably due to insufficient sequences overlap). We retrieved 167 UP1 and 167 SO1 as well as 1,656 UP2 and 1,203 SO2 clusters. After cleaning, our final dataset for codeml analysis comprised 160 UP1, 1,512 UP2, 167 SO1, and 1,203 SO2 clusters for the codeml analysis (Table 1).

As alignment errors can create false positives in the detection of positive selection footprints, each cluster suggested to be under positive selection was again checked both automatically – using muscle [72] and trimAL [73] for creating and cleaning alignments (muscle-trimAL method; see Additional file 1) – and manually for alignment errors. We found that our initial alignment and cleaning procedure using PRANK [71] and GUIDANCE [42] is superior to the muscle-trimAL method. Manual curation, however, remains essential to avoid false positives (Additional file 1).

Detecting codons under positive selection

We used codeml site model implemented in the PAML4 software [74] to infer positive selection on codons under several substitution models. For these analyses, we extensively relied on the egglib package [75] to implement the following pipeline: First, for every alignment the maximum likelihood phylogeny was inferred at the nucleotide level using PhyML 3.0 [76] under the GTR-Γ substitution model (Figure 1c). Second, different codeml site models were run (Figure 1d). The nearly neutral models (M1a and M8a) assume codons to evolve either neutrally or under purifying selection whereas the positive selection models (M2a and M8) assume positive selection acting on some codons. Third, likelihood ratio tests (LRTs) were performed using R [77] to compare nearly neutral and positive selection models and hence to detect clusters for which models including positive selection are significantly more likely than models that do not. We corrected for multiple testing using a Bonferroni correction. In clusters identified to have evolved under positive selection, Bayes empirical Bayes was used to calculate the posterior probabilities at each codon and detect those under positive selection (*i.e.* those with a posterior probability of $\omega > 1$ strictly above 95%). All alignments detected to be under positive selection at the codon level were curated manually for potential alignment errors. More details on the estimated omega for each cluster with codons under positive selection, position of every codon under positive selection, and results of the LRT for those clusters can be found in Additional file 3. All cleaned alignments containing codons under positive selection are provided in Additional file 4.

Assessing d_N/d_S at branches

For inferring ω on branches, the alignments and the corresponding phylogenies were used as input for mapNH [78,79]. Unlike the branch-site model in codeml, this method does not require to define branches under selection *a priori* [78]. mapNH performs substitution mapping before clustering branches according to their underlying substitution processes (Figure 1d). The ω of each branch was then calculated as followed:

$$\omega = \frac{nbNS/NSsites}{nbS/Ssites}$$

using *nbNS* (number of non-synonymous mutations) and *nbS* (number of synonymous mutations) estimations provided by mapNH whereas *NSsites* (number of non-synonymous sites) and *Ssites* (number of synonymous sites) were computed by codeml during the site model analysis. We preferably used the *NSsites* and *Ssites* provided by codeml since they benefit from the maximum likelihood estimation of the transition/transversion ratio done by codeml for each alignment. Finally, note that ω was estimated only for clusters with at least one synonymous and one non-synonymous mutation. After clusters with no mutation were removed for the mapNH analysis, 154 UP1, 1,435 UP2, 167 SO1, and 1,203 SO2 clusters remained (Table 1). Branches containing no substitutions were also removed, leaving us with 1,257 UP1, 14,326 UP2, 1,807 SO1, and 13,374 SO2 branches for the final analysis (Table 1).

Determining effects of time and polymorphism

SO and UP clusters are different by definition. First, the divergence times between sequences are not expected to be the same. Specifically, divergence in a given SO cluster should range between minimum and maximum divergence time of the species included in this cluster. Divergence in UP clusters should range from null (for very recent duplications) to the last speciation event. It has been shown that d_N/d_S-based tests are strongly influenced by d_S [59]. To test whether our results are biased due to divergence discrepancies between UP and SO, we sorted the ω value of each branch by their synonymous substitution rate (d_S). Second, in the UP dataset some duplications could have occurred very recently. It is likely that some differences between those young paralogs are still segregating in populations and should therefore be considered as polymorphism instead of divergence. Inferring selection using d_N/d_S in such a scenario has been shown to be incorrect [60]. To rule out effects of polymorphism on UP clusters, we excluded external branches from the dataset and compared the remaining internal branches to the SO dataset. To test if ω differs significantly between types of clusters, we performed a Mann–Whitney test using R

[77]. When ω is analyzed according to d_S, Mann–Whitney tests were performed in a sliding window of 0.01 d_S. The calculation was done when a window contained at least 100 values by group studied.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Additional files

Additional file 1: Extended Materials and Methods and extended Results. This includes: **Table S1:** Clusters initially under selection and number and percentage of clusters removed after manual inspection and after applying the muscle-trimAL pipeline for Level 1 the dataset. **Table S2:** Clusters containing codons under positive selection according to the codeml site model before and after manual curation in the whole dataset. **Table S3:** Clusters containing codons under positive selection according to the codeml site model before and after manual curation in the dataset containing clusters from GreenPhyl trees with only one UP and/or SO cluster. **Table S4:** Clusters containing codons under positive selection according to the codeml site model before and after manual curation in the dataset containing only clusters from trees harboring several clusters. **Table S5:** Results of the mapNH analysis for the different datasets. **Figure S1:** Phylogeny of a subset of plant species of the GreenPhylDB. **Figure S2:** Overview of the different sub-datasets analyzed.

Additional file 2: Excel spreadsheet containing GreenPhyl gene families found to be under positive selection with the codeml site model.

Additional file 3: Excel spreadsheet that contains the estimated omega for each cluster with codons under selection, results of the LRT for those clusters, and the position of every manually curated codon under positive selection in the provided alignments.

Additional file 4: Contains the cleaned alignments of the clusters with codons under selection.

Abbreviations

WGD: Whole genome duplication; SSD: Short scale genomic duplication; TDG: Tandemly duplicated gene; LSE: Lineage specific expansion; UP: Ultraparalog; SO: Superortholog; LRR: Leucine-rich repeat; RLK: Receptor-like kinase; FBP: F-box protein.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

IF, JD, VR, JFD, and NC designed the study; IF, JD, and JFD developed the pipeline; IF, JD, SG, and JFD performed the data analysis and statistics; IF drafted the manuscript with the help of JD, VR, SG, JFD, and NC. All authors read and approved of the final manuscript.

Acknowledgements

We are grateful to Mathieu Rouard for providing the GreenPhylDB data and the SouthGreen bioinformatics platform (http://www.southgreen.fr/) which was essential for our analysis. We also would like to thank two anonymous reviewers for their constructive comments on the manuscript. IF was granted by ARCAD (Agropolis Resource Center for Crop Conservation, Adaptation and Diversity), a flagship project of the Agropolis Fondation, and JD by scientific council of Montpellier SupAgro. This work was partially funded by the French *Agence Nationale de la Recherche*: ANR-11-BSV7-013-03, TRANS and by the French *Agence Nationale de la Recherche Investissements d'avenir/ Bioinformatique*: ANR-10-BINF-01-02, Ancestrome.

Author details

[1]INRA, UMR 1334 AGAP, 2 Place Pierre Viala, 34060 Montpellier, France. [2]IRD, UMR 232 DIADE, 911 Avenue Agropolis, 34394 Montpellier, France. [3]Montpellier SupAgro, UMR 1334 AGAP, 2 Place Pierre Viala, 34060 Montpellier, France. [4]Université Montpellier II, Institut des Sciences de l'Evolution CC64, Place Eugène Bataillon, 34095 Montpellier, France. [5]CIRAD, UMR 1334 AGAP, Avenue Agropolis, 34398 Montpellier, France. [6]Present Address: Department of Medical Biochemistry, Microbiology, Genomics, Uppsala University, Husargatan 3, 75123 Uppsala, Sweden.

Received: 10 March 2014 Accepted: 23 May 2014
Published: 31 May 2014

References

1. Nei M, Rooney AP: **Concerted and birth-and-death evolution of multigene families.** *Annu Rev Genet* 2005, **39:**121–152.
2. Kejnovsky E, Leitch IJ, Leitch AR: **Contrasting evolutionary dynamics between angiosperm and mammalian genomes.** *Trends Ecol Evol* 2009, **24:**572–582.
3. Leitch AR, Leitch IJ: **Ecological and genetic factors linked to contrasting genome dynamics in seed plants.** *New Phytol* 2012, **194:**629–646.
4. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449:**463–467.
5. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengelle J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, McKain MR, Leebens-Mack J, Burgess D, Freeling M, Mbeguie AMD, Chabannes M, Wicker T, *et al*: **The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants.** *Nature* 2012, **488:**213–217.
6. The Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution.** *Nature* 2012, **485:**635–641.
7. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K: **The flowering world: a tale of duplications.** *Trends Plant Sci* 2009, **14:**680–688.
8. Initiative TAG: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796–815.
9. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436:**793–800.
10. Rizzon C, Ponger L, Gaut BS: **Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice.** *PLoS Comput Biol* 2006, **2:**e115.
11. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH: **Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli.** *Plant Physiol* 2008, **148:**993–1003.
12. Jiang S-Y, Gonzalez JM, Ramachandran S: **Comparative genomic and transcriptomic analysis of tandemly and segmentally duplicated genes in rice.** *PLoS One* 2013, **8:**e63551.
13. Kondrashov FA: **Gene duplication as a mechanism of genomic adaptation to a changing environment.** *Proc R Soc Lond B* 2012, **279:**5048–5057.
14. Lynch M: **Genomic expansion by gene duplication**. In *The origins of genome architecture.* Edited by. Sunderland, MA, USA: Sinauer Associates, Inc; 2007.
15. Moore RC, Purugganan MD: **The evolutionary dynamics of plant duplicate genes.** *Curr Opin Plant Biol* 2005, **8:**122–128.
16. Hahn MW: **Distinguishing among evolutionary models for the maintenance of gene duplicates.** *J Hered* 2009, **100:**605–617.
17. Innan H: **Population genetic models of duplicated genes.** *Genetica* 2009, **137:**19–37.
18. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nature Rev Genet* 2010, **11:**97–108.
19. Francino MP: **An adaptive radiation model for the origin of new gene functions.** *Nat Genet* 2005, **37:**573–577.
20. Piatigorsky J, Wistow G: **The recruitment of crystallins: new functions precede gene duplication.** *Science* 1991, **252:**1078–1079.
21. Hughes AL: **The evolution of functionally novel proteins after gene duplication.** *Proc R Soc Lond B* 1994, **256:**119–124.

22. Des Marais DL, Rausher MD: **Escape from adaptive conflict after duplication in an anthocyanin pathway gene.** *Nature* 2008, **454:**762–765.
23. Otto SP, Whitton J: **Polyploid incidence and evolution.** *Annu Rev Genet* 2000, **34:**401–437.
24. Mano S, Innan H: **The evolutionary rate of duplicated genes under concerted evolution.** *Genetics* 2008, **180:**493–505.
25. Doyon J-P, Ranwez V, Daubin V, Berry V: **Models, algorithms and programs for phylogeny reconciliation.** *Brief Bioinform* 2011, **12:**392–400.
26. van der Heijden RTJM, Snel B, van Noort V, Huynen MA: **Orthology prediction at scalable resolution by phylogenetic tree analysis.** *BMC Bioinformatics* 2007, **8:**83.
27. Storm CEV, Sonnhammer ELL: **Orthology prediction at scalable resolution by phylogenetic tree analysis.** *Bioinformatics* 2002, **18:**92–99.
28. Yang X, Kalluri UC, Jawdy S, Gunter LE, Yin T, Tschaplinski TJ, Weston DJ, Ranjan P, Tuskan GA: **The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants.** *Plant Physiol* 2008, **148:**1189–1200.
29. Aguilar-Hernández V, Aguilar-Henonin L, Guzmán P: **Diversity in the architecture of ATLs, a family of plant ubiquitin-ligases, leads to recognition and targeting of substrates in different cellular environments.** *PLoS One* 2011, **6:**e23934.
30. Hua Z, Zou C, Shiu SH, Vierstra RD: **Phylogenetic comparison of *F-Box* (*FBX*) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift.** *PLoS One* 2011, **6:**e16219.
31. Tuominen LK, Johnson VE, Tsai C-J: **Differential phylogenetic expansions in BAHD acyltransferases across five angiosperm taxa and evidence of divergent expression among *Populus* paralogues.** *BMC Genomics* 2011, **12:**236.
32. Yonekura-Sakakibara K, Hanada K: **An evolutionary view of functional diversity in family 1 glycosyltransferases.** *Plant J* 2011, **66:**182–193.
33. Martinez-Castilla LP, Alvarez-Buylla ER: **Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny.** *Proc Natl Acad Sci USA* 2003, **100:**13407–13412.
34. Johnson DA, Thomas MA: **The monosaccharide transporter gene family in *Arabidopsis* and rice: a history of duplications, adaptive evolution, and functional divergence.** *Mol Biol Evol* 2007, **24:**2412–2423.
35. Xue Z, Duan L, Liu D, Guo J, Ge S, Dicks J, ÓMáille P, Osbourn A, Qi X: **Divergent evolution of oxidosqualene cyclases in plants.** *New Phytol* 2012, **193:**1022–1038.
36. Smith SD, Wang S, Rausher MD: **Functional evolution of an anthocyanin pathway enzyme during a flower color transition.** *Mol Biol Evol* 2013, **30:**602–612.
37. Yang Z, Wang Y, Zhou Y, Gao Q, Zhang E, Zhu L, Hu Y, Xu C: **Evolution of land plant genes encoding L-Ala-D/L-Glu epimerases (AEEs) via horizontal gene transfer and positive selection.** *BMC Plant Biol* 2013, **13:**34.
38. Wong WS, Yang Z, Goldman N, Nielsen R: **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites.** *Genetics* 2004, **168:**1041–1051.
39. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution.** *Mol Biol Evol* 2001, **18:**1585–1592.
40. Zhang J, Nielsen R, Yang Z: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.** *Mol Biol Evol* 2005, **22:**2472–2479.
41. Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Périn C, Conte MG: **GreenPhylDB v2.0: comparative and functional genomics in plants.** *Nucleic Acids Res* 2011, **39:**D1095–D1102.
42. Penn O, Privman E, Landan G, Graur D, Pupko T: **An alignment confidence score capturing robustness to guide tree uncertainty.** *Mol Biol Evol* 2010, **27:**1759–1767.
43. Flagel LE, Wendel JF: **Gene duplication and evolutionary novelty in plants.** *New Phytol* 2009, **183:**557–564.
44. Hughes AL: **Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level.** *Heredity (Edinb)* 2007, **99:**364–373.
45. Spofford JB: **Heterosis and evolution of duplications.** *Amer Nat* 1969, **103:**407–432.
46. Lynch M, O'Hely M, Walsh B, Force A: **The probability of preservation of a newly arisen gene duplicate.** *Genetics* 2001, **159:**1789–1804.
47. Proulx SR, Phillips PC: **Allelic divergence precedes and promotes gene duplication.** *Evolution* 2006, **60:**881–892.
48. De Bodt S, Maere S, Van de Peer Y: **Genome duplication and the origin of angiosperms.** *Trends Ecol Evol* 2005, **20:**591–597.
49. Amborella Genome Project: **The *Amborella* genome and the evolution of flowering plants.** *Science* 2013, **342:**. doi: 10.1126/science.1241089.
50. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW: **Ancestral polyploidy in seed plants and angiosperms.** *Nature* 2011, **473:**97–100.
51. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR, Weigel D: **Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*.** *Science* 2007, **317:**338–342.
52. Schumann N, Navarro-Quezada A, Ullrich K, Kuhl C, Quint M: **Molecular evolution and selection patterns of plant F-box proteins with C-terminal kelch repeats.** *Plant Physiol* 2011, **155:**835–850.
53. Navarro-Quezada A, Schumann N, Quint M: **Plant F-box protein evolution is determined by lineage-specific timing of major gene family expansion waves.** *PLoS One* 2013, **8:**e68672.
54. Dievart A, Gilbert N, Droc G, Attard A, Gourgues M, Guiderdoni E, Perin C: **Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes.** *BMC Evol Biol* 2011, **11:**367.
55. Tang P, Zhang Y, Sun X, Tian D, Yang S, Ding J: **Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species.** *Plant Sci* 2010, **179:**399–406.
56. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3.** *Mol Biol Evol* 2013, **30:**1987–1997.
57. Fletcher W, Yang Z: **The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection.** *Mol Biol Evol* 2010, **27:**2257–2267.
58. Jordan G, Goldman N: **The effects of alignment error and alignment filtering on the sitewise detection of positive selection.** *Mol Biol Evol* 2012, **29:**1125–1139.
59. Wolf JB, Kunstner A, Nam K, Jakobsson M, Ellegren H: **Nonlinear dynamics of nonsynonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection.** *Genome Biol Evol* 2009, **1:**308–319.
60. Kryazhimskiy S, Plotkin JB: **The population genetics of d_N/d_S.** *PLoS Genet* 2008, **4:**e1000304.
61. Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Paterson AH: **Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms.** *PLoS One* 2011, **6:**e28150.
62. Wang Y: **Locally duplicated ohnologs evolve faster than nonlocally duplicated ohnologs in Arabidopsis and rice.** *Genome Biol Evol* 2013, **5:**362–369.
63. Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH: **Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana.** *PLoS Genet* 2009, **5:**e1000581.
64. The International Brachypodium Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon*.** *Nature* 2010, **463:**763–768.
65. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112–1115.
66. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, *et al*: **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature* 2009, **457:**551–556.
67. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313:**1596–1604.

68. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, *et al*: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463:**178–183.
69. Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, *et al*: **The *Medicago* genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480:**520–524.
70. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G: **Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases.** *Bioinformatics* 2005, **21:**2596–2603.
71. Löytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions.** *Proc Natl Acad Sci USA* 2005, **102:**10557–10562.
72. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.
73. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25:**1972–1973.
74. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.
75. De Mita S, Siol M: **EggLib: processing, analysis and simulation tools for population genetics and genomics.** *BMC Genet* 2012, **13:**27.
76. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307–321.
77. R Development Core Team: *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing; 2012.
78. Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B: **Efficient selection of branch-specific models of sequence evolution.** *Mol Biol Evol* 2012, **29:**1861–1874.
79. Romiguier J, Figuet E, Galtier N, Douzery EJ, Boussau B, Dutheil JY, Ranwez V: **Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping.** *PLoS One* 2012, **7:**e33852.

doi:10.1186/1471-2229-14-151
Cite this article as: Fischer *et al.*: **Impact of recurrent gene duplication on adaptation of plant genomes.** *BMC Plant Biology* 2014 **14**:151.

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Contrasted patterns of selective pressure in three recent paralogous gene pairs in the *Medicago* genus (L.)

Joan Ho-Huu[1], Joëlle Ronfort[1], Stéphane De Mita[1,2], Thomas Bataillon[3], Isabelle Hochu[1], Audrey Weber[1] and Nathalie Chantret[1*]

Abstract

Background: Gene duplications are a molecular mechanism potentially mediating generation of functional novelty. However, the probabilities of maintenance and functional divergence of duplicated genes are shaped by selective pressures acting on gene copies immediately after the duplication event. The ratio of non-synonymous to synonymous substitution rates in protein-coding sequences provides a means to investigate selective pressures based on genic sequences. Three molecular signatures can reveal early stages of functional divergence between gene copies: change in the level of purifying selection between paralogous genes, occurrence of positive selection, and transient relaxed purifying selection following gene duplication. We studied three pairs of genes that are known to be involved in an interaction with symbiotic bacteria and were recently duplicated in the history of the *Medicago* genus (Fabaceae). We sequenced two pairs of polygalacturonase genes (*Pg11*-*Pg3* and *Pg11a*-*Pg11c*) and one pair of auxine transporter-like genes (*Lax2*-*Lax4*) in 17 species belonging to the *Medicago* genus, and sought for molecular signatures of differentiation between copies.

Results: Selective histories revealed by these three signatures of molecular differentiation were found to be markedly different between each pair of paralogs. We found sites under positive selection in the *Pg11* paralogs while *Pg3* has mainly evolved under purifying selection. The most recent paralogs examined *Pg11a* and *Pg11c,* are both undergoing positive selection and might be acquiring new functions. *Lax2* and *Lax4* paralogs are both under strong purifying selection, but still underwent a temporary relaxation of purifying selection immediately after duplication.

Conclusions: This study illustrates the variety of selective pressures undergone by duplicated genes and the effect of age of the duplication. We found that relaxation of selective constraints immediately after duplication might promote adaptive divergence.

Keywords: Duplication, Medicago, Neofunctionalization, Subfunctionalization, Paralogs evolution

Background

Gene duplications have long been hypothesized to be drivers of genome and gene function evolution [1]. Recently, availability of large-scale sequence data, and especially entire genome sequences, has brought significant support to this view [2,3]. In plants, duplications appear to be frequent and most lineages studied up to now have been affected by whole-genome duplication events (polyploidy) and/or segmental duplications [4-10].

Starting with Ohno, a range of models has been proposed to predict the fates of paralogous gene pairs resulting from duplications. These models can be categorized by their assumptions: they can be either neutral or involving natural selection, and can consider the early stage of duplication, *i.e.* when the duplication is not yet fixed in the species or start with the assumption that the gene duplication has just been fixed (recently reviewed in [11]).

* Correspondence: Nathalie.Chantret@supagro.inra.fr
[1]INRA - Institut National de la Recherche Agronomique, UMR AGAP, Montpellier 34060, France
Full list of author information is available at the end of the article

Immediately after the gene duplication event, the two copies are assumed to be identical and therefore functionally redundant. At this stage, there should be no selective pressure against any loss-of-function mutation affecting either copy. As a result, it is believed that most instances of gene duplications will eventually result in the loss of one of the copies (pseudogenization or nonfunctionalization). However, the relaxation of purifying selection (due to the initial redundancy) may allow some amount of divergence and occasionally can let one copy acquire a new function and be subsequently maintained by natural selection (neofunctionalization). This scenario is essential for the creative role of duplication envisioned by Ohno [1]. Force *et al.* [12] suggested that the presence of two redundant genes may drive the fixation of complementary degenerative mutations in both of copies, with higher probability in gene regulatory regions. At the end of this process, both gene copies are required to perform the set of functions originally performed by a single gene (subfunctionalization). These two scenarios are not mutually exclusive and may act jointly [13]. Besides these models, the maintenance of functionally redundant copies (without functional divergence) could be adaptive under specific circumstances, either through dosage effect or as a means of genetic robustness against deleterious mutations [14-16] and therefore also explain the fixation of duplications in species [11].

Functional analyses have been performed in order to determine the relative importance or the interaction between these different models. The occurrence and the characteristics of functional divergence of paralogous genes can be addressed either through the regulatory or protein-coding sequence angle.

Whole-genome expression profiles revealed divergent expression patterns between paralogous gene pairs, providing indirect evidence for subfonctionalization and/or neofunctionalization [17]. Similar conclusions were also drawn from studies of polyploid species for which duplicated genes were instantly fixed in the species founder individual [18-20]. More specific and detailed functional analyses revealed several cases of paralogs undergoing neofunctionalization or subfunctionalization [21,22].

Beside differences in gene expression, rates of molecular evolution can be used to qualify the constraints experienced by genes. In particular, contrasting the rate of protein-changing (non-synonymous) substitution (dN) and the rate of silent (synonymous) substitution (dS) at the nucleotide level allows qualifying the type of selection acting on individual gene copies after a duplication event. The intensity of purifying selection is often estimated through the ratio $\omega = dN/dS$. Values of $\omega < 1$ are interpreted as evidence for purifying selection (the lower ω, the stronger purifying selection). Following pseudogenization, $\omega = 1$ is expected (no constraint). Last, amino acid sites exhibiting $\omega > 1$ are likely directly targeted by positive selection. As an example, the evolutionary fate of ten genes recently duplicated by retrotransposition in mice was studied by contrasting synonymous and non-synonymous rates [23]. Gene duplications have been the subject of many functional and molecular studies in plants [24,25], but here we aimed at analysing specifically the selective constraints exerted on duplicated genes through analysis of their rates of substitution. In order to shed light to the temporal variation of selective constraints acting on duplicated genes following their duplication, we focused on the evolution of fairly recent duplicated genes at a time scale appropriate for coding sequence evolution rates analysis. Such study can provide insight about the relative role of relaxation of purifying selection and positive selection in the fate of duplicated genes.

We investigated rates of molecular evolution of three duplicated gene pairs in the genus *Medicago* (L.), therefore maximizing the amount of available phylogenetic signal. We selected gene pairs involved directly or indirectly in the symbiotic interaction between legumes and nitrogen-fixing bacteria (rhizobia). The first genes code for polygalacturonases, which are enzymes involved in the degradation of polysaccharides. One member (*Pg11*) is involved in pollen tube elongation and the other (*Pg3*) in the tip growth of the infection threads during the establishment of the symbiosis with nitrogen-fixing bacteria *Sinorhizobium* sp [26]. The second genes are *Lax* (Like-*Aux1*). They are auxin efflux carriers and play an important role in auxin-controlled processes such as tissue growth and in particular development of nodules.

Mutualistic host-symbiont interactions present the interest of combining several features we can expect will promot fast evolution. Mutualisms are often based on nutrient exchanges and involve strong selective pressures, since both costs and benefits are important. The interaction with a biotic partner can cause shifting selective optima, especially if there are conflicts of interest. Finally, in contrast with host-pathogen interactions, mutualisms can involve the evolution of novel structures by both partners. The legume-rhizobium symbiosis evolved relatively recently, around 60 million years ago, culminating with the emergence of a specific organ, the root nodule [27]. Therefore, the genes underlying rhizobial symbiosis in legumes are likely to record the signatures of past selective pressures caused by the emergence and diversification of symbiosis as well as pressures linked to their current function. Due to a whole-genome duplication event that occurred approximately 58 Myr ago [28], legumes are therefore a good model to examine the changes of selective pressures over time for duplicated genes.

Rates of molecular evolution of paralogous gene copies (hereafter paralogs) should be studied preferably in a variety of species to have enough power to inner substitution rates. Moreover paralogs should be characterized in a set of extant species that have diverged after the ancestral gene duplication. In spite the growing availability of full genome sequences, plant model species are usually not related enough to allow for analysis of divergence at the nucleotide level. In the case of the Fabaceae family, three species have been sequence (*Medicago truncatula*, *Lotus japonicus* and *Glycine max*), but their divergence times would represent a time scale of 50–60 million years [29]. Moreover, more taxa are needed for contrasting early and late selective pressures. We resequenced three pairs of relatively recently duplicated genes in 16 other species of the *Medicago* genus (in addition to *M. truncatula*). We chose duplicated genes that (i) are recent enough so that the signatures of evolution post-duplication are still detectable, (ii) predate the speciation events within the *Medicago* genus, so that each copy is found within all species and (iii) contain at least one gene demonstrated or strongly suspected to be involved in the legume-specific symbiotic interaction with nitrogen-fixing rhizobium bacteria.

Results

Sequencing *Pg11a*, *Pg11c*, *Lax2* and *Lax4*

Depending of the gene, a successful amplification was obtained for a total of 10 to 17 species. The resulting sequence alignments had a length of 729 bp for *Pg* genes and 798 bp for *Lax* genes. We excluded sequences that did not encode a complete protein (due to frame shift or nonsense mutations) because they might represent pseudogenes and affect our estimates of rates of molecular evolution in functional paralogs. Accession numbers of sequences deposited in GenBank are from JN635641 to JN635687. Already available sequences GenBank accession numbers are AJ620946, AY115843 and AY115844 (for *M. truncatula* genes *Pg3*, *Lax2* and *Lax4* respectively), HQ737838, HQ736585 and HQ736701 (for *M. tornata* genes *Pg3*, *Lax2* and *Lax4* respectively). Details about the sequences obtained as well as GenBank accession numbers are given in Additional file 1.

Phylogeny of *Pg* and *Lax* genes

The phylogeny of *Lax* and *Pg* paralogs were reconstructed using maximum likelihood and are presented in Figure 1a and Figure 1b respectively. The topologies obtained for the three paralog pairs *Lax2*/*Lax4*, *Pg3*/*Pg11* and *Pg11a*/*Pg11c* confirmed the occurrence of three duplication events predating the divergence between the 17 species we included from the *Medicago* genus. Moreover, the branches leading to each paralog clade containing the sequences of the same gene amplified from different species are well supported. Bootstrap values for the branches leading to the *Lax2*, *Lax4*, *Pg3* and *Pg11* clades are equal to 100. Within the *Pg11* clade, bootstrap values obtained for the branches leading to the *Pg11a* and *Pg11c* clades are 95 and 93 respectively. However, within the *Pg11a* and *Pg11c* clades, several inconsistencies were observed in the phylogenies: sequences obtained with the *Pg11c* copy specific primers for *M. carstiensis*, *M. ruthenica* and *M sauvagei* were placed in the *Pg11a* clade, and conversely, sequences obtained with the *Pg11a* copy specific primers for *M. tornata* were placed in the *Pg11c* clade (highlighted in grey in Figure 1b). There are several explanations for such inconsistencies: erroneous amplification (for example chimeric amplification), the amplification of a third copy resulting from an independent duplication, or genic conversion between paralogs. In order to avoid erroneous interpretations, we did not consider these four sequences further in our analysis.

The species phylogeny deduced from the data was not completely congruent between different paralogs and with the species phylogeny described in the literature [30,31]. However, within clades regrouping sequences of a same gene in the different species, branches are not well supported (Figure 1), indicating a poor phylogenetic resolution. Only three groups of species were grouped with high support, irrespective of the gene analysed. The first includes *M. tornata*, *M. truncatula*, *M. tricycla* and *M. littoralis*, the second*M. rigidula* and *M. rigiduloïdes* and the third *M. sauvagei* and *M. laciniata*. The *Medicago* genus evolved through a large number of speciation events in a short time span, and as a result, the resolution of phylogenetic relationships between *Medicago* species is difficult. Furthermore, incongruences may be observed between gene and species trees due to incomplete lineage sorting [32]. For each paralog set, we used the best fitting phylogenetic tree. We repeated the analyses of selective constraints for each gene pair using either the best topology found for the genes considered or a tree topology from the literature [30]. Results were very similar and conclusions were not affected. Consequently, only results obtained using the phylogeny from our data are presented.

Analysis of selective pressures along trees: testing for an "age" and a "paralog" effect

Comparing models with different constraints on the value of ω among branches of the tree allows testing evolutionary hypotheses (Figure 2a). The comparisons of M_A versus M_0 and M_{PA} versus M_P test the "age effect" by contrasting early branches (when the duplication was young) and later branches. Similarly, the comparisons of M_P versus M_0 and M_{PA} versus M_A test a "paralog effect" by examining the divergence between the two copies. Results of these tests are

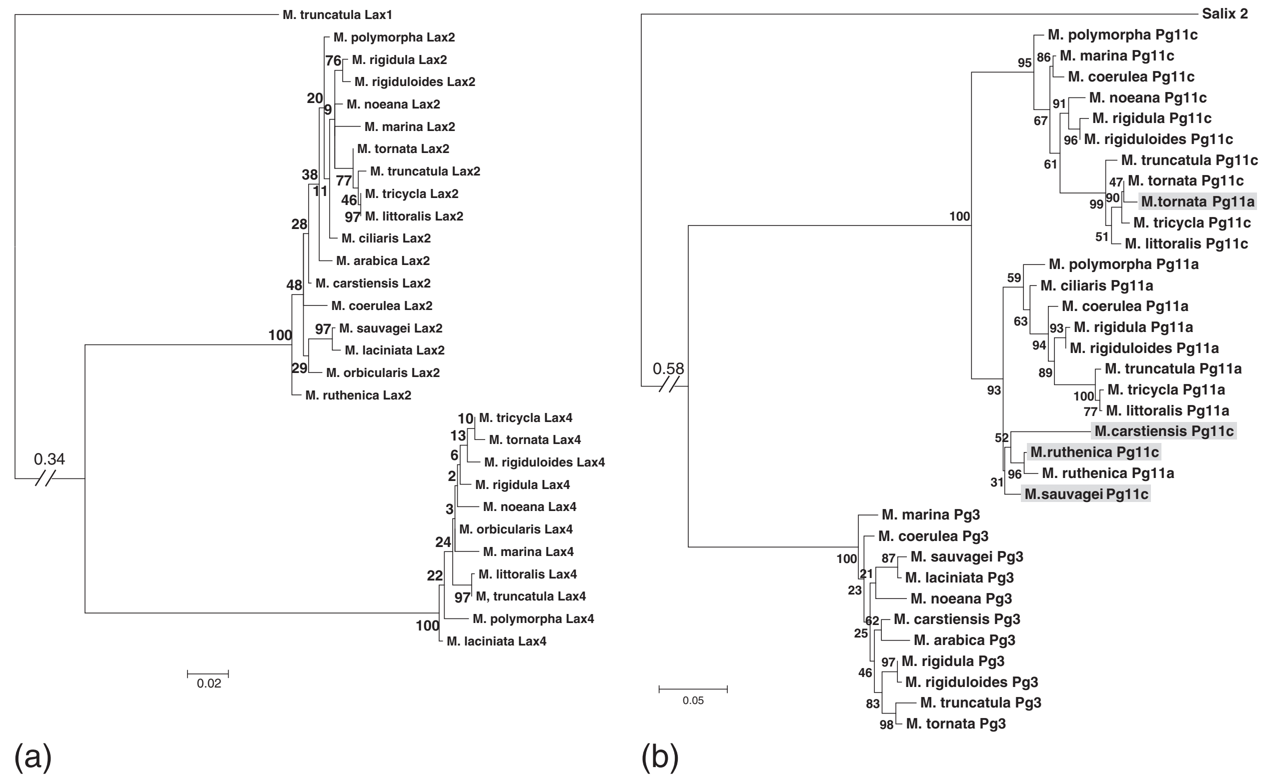


Figure 1 Phylogenetic trees for *Lax* and *Pg* genes. Phylogenetic trees obtained for *Lax2* and *Lax4* (**a**) and for *Pg3*, *Pg11a* and *Pg11c* genes (**b**). Bootstrap values (% of 100 re-sampled data set) are indicated for each branch. For presentation convenience, the branch leading to the outgroup was shortened (as indicated by an inclined double segment). Sequences underlined in light grey represent instances were a paralog is placed in the wrong clade when considering the copy specific primers used for its amplification.

presented in Table 1 along with maximum-likelihood estimates of ω parameters of each model. When the *Pg11*/*Pg3* paralogs pair was analysed, both the sequences of *Pg11a* and *Pg11c* were considered for *Pg11*. For example, for testing the M_P model, both branches leading to *Pg11a* and *Pg11c* were considered for *Pg11*. For testing the M_A model, the branch between the node corresponding to the duplication between *Pg11* and *Pg3* and the node corresponding to the duplication between *Pg11a* and *Pg11c* (*i.e.* the ancestral *Pg11* gene) was considered as the late branch. Thus, any effects related to the duplication between *Pg11a* and *Pg11c* is considered only in the *Pg11a*/*Pg11c* paralogs pair analysis.

Interestingly, the three paralogs pairs exhibited contrasted results. The *Lax2*/*Lax4* paralogs shows evidence for an age effect as shown by both a better fit of model M_A relative to M_0 (LRT = 13.0, $p = 0.00031$) and M_{PA} versus M_P (LRT = 13.8, $p = 0.001$) tests. We observed a marked increase of ω in early branches (ω = 0.14 compared with ω = 0.05 for late branches in M_A). No significant paralog effect was detected and both paralogs *Lax2* and *Lax4* seem to be evolving under purifying selection (ω = 0.08 in M_0, ω = 0.07 and ω = 0.10 for *Lax2* and *Lax4* respectively in M_P model).

For the second paralog pair, *Pg11*/*Pg3*, both age (M_A vs. M_0, LRT = 4.84, p = 0.03) and paralog effects (M_P vs. M_0, LRT = 4.28, p = 0.04) are detected, but effects were weaker and marginally significant. The full model (M_{PA}) did not provide a better fit relative to either the age or the copy models. The M_A model showed an increase of ω in late branches (ω = 0.38 as compared to 0.22 in early branches) and an increase of ω in *Pg11* (ω = 0.41 as compared to 0.25 in *Pg3*).

For the third and most recent pair of paralogous genes, *Pg11a*/*Pg11c*, no test was significant, suggesting that neither age nor paralog effects is playing a role or that the extent of nucleotide differences are too small for codon based models to have any power to detect heterogeneity in ω. The analysis shows that overall ω is markedly higher than in the other considered paralog pairs (ω = 0.44 for M_0).

Analysis of selective pressures along genes: testing for positive selection

In order to investigate how ω varies along genes and in particular if positive selection signatures occurred we used models in which ω is allowed to vary among sites on each gene (Figure 2b). As in analysis of selective pressures along

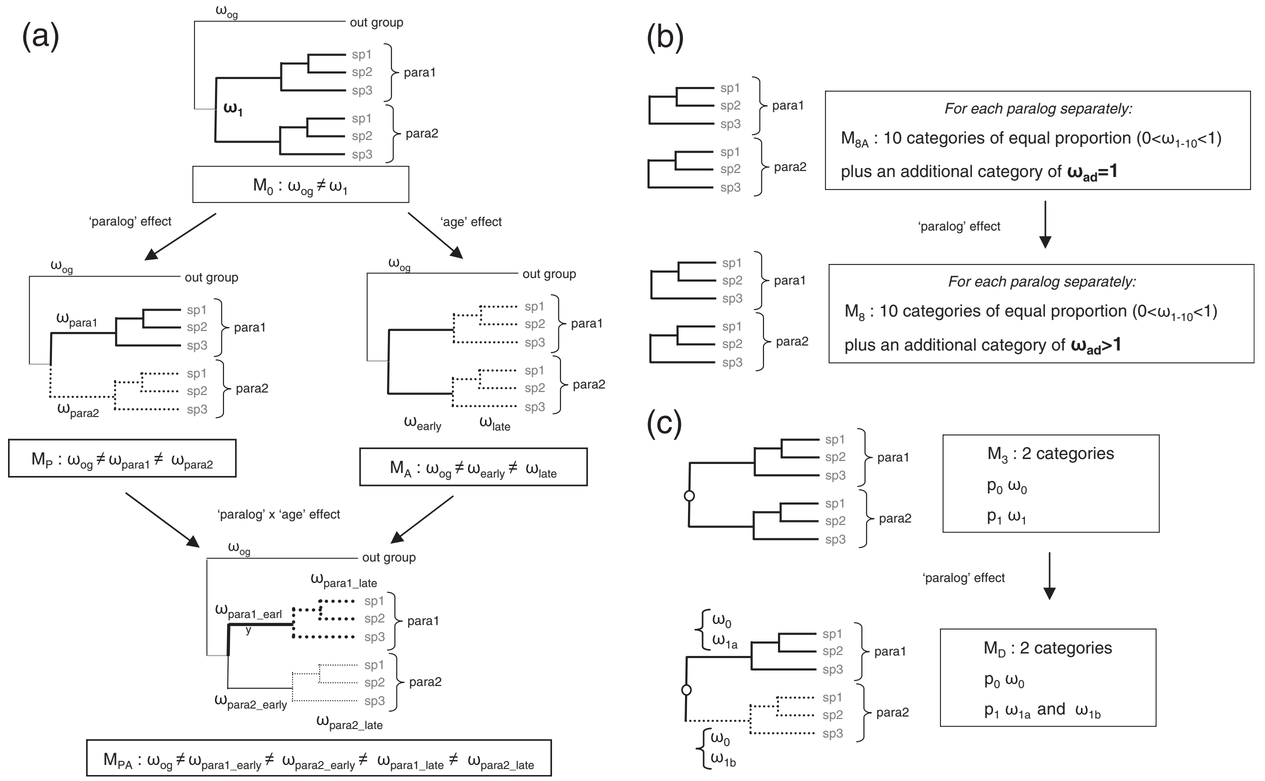

(a)
ω_og
out group
sp1
sp2
sp3
para1
ω_1
para2
M_0 : ω_og ≠ ω_1
'paralog' effect
'age' effect
ω_para1
ω_para2
M_P : ω_og ≠ ω_para1 ≠ ω_para2
ω_early
ω_late
M_A : ω_og ≠ ω_early ≠ ω_late
'paralog' x 'age' effect
ω_para1_late
ω_para1_early
ω_para2_early
ω_para2_late
M_PA : ω_og ≠ ω_para1_early ≠ ω_para2_early ≠ ω_para1_late ≠ ω_para2_late
(b)
For each paralog separately:
M_8A : 10 categories of equal proportion (0<ω_1-10<1)
plus an additional category of ω_ad=1
'paralog' effect
For each paralog separately:
M_8 : 10 categories of equal proportion (0<ω_1-10<1)
plus an additional category of ω_ad>1
(c)
M_3 : 2 categories
p_0 ω_0
p_1 ω_1
'paralog' effect
ω_0
ω_1a
ω_0
ω_1b
M_D : 2 categories
p_0 ω_0
p_1 ω_1a and ω_1b

Figure 2 Schematic representation of the codon models used. (**a**) Models allowing dN/dS variation along lineages. Arrows indicate the questions addressed by the comparison between models (in (**a**), the arrows correspond to hierachical relationships). (**b**) Models allowing dN/dS variation along the gene. In M_{8A}, ω follows a β distribution discretized into 10 categories of similar frequency ($0 < \omega_{1-10} < 1$) and an additional category of ω is fixed at 1 ($\omega_{ad} = 1$, accounting for neutral sites); M_8 differs from M_{8A} only by the additional category of ω which is constraint to be superior to 1 ($\omega_{ad} > 1$), to account for sites under positive selection. (**c**) Phylogenetic trees harbours two clades, one for each paralog (the outgroup is not represented here). Models are either specifying identical categories of dN/dS in both clades (M_3), or allowing one category to take a different value in each clade (M_D).

trees, when *Pg11* gene is analysed, both sequences of *Pg11a* and *Pg11c* were considered. Since positive selection likely targets only a few amino acid positions, branch models used previously typically lack statistical power to detect positive selection as, in the branch model, ω is averaged over all the amino acid sites of the gene.

We compared the fit of models M_{8A} and M_8. The likelihood ratio test of M_8 against M_{8A} is a conservative test for positive selection, since the M_{8A} model can account for an excess of neutral sites. No sites under positive selection were found for *Lax2*, *Lax4* and *Pg3*. However the statistical test of M_8 against M_{8A} was significant for *Pg11*, *Pg11a* and *Pg11c* ($p = 1.4\ 10^{-6}$, $9.99\ 10^{-3}$ and $1.26\ 10^{-4}$ respectively) (Table 2), showing that positive selection targeted both copies of *Pg11*. The fitted ω values suggest that positive selection was stronger for *Pg11a* ($\omega = 11.61$ at positively selected sites) than for *Pg11c* ($\omega = 4.45$) but affected fewer sites (frequency of 0.02, equivalent to 1 site, for *Pg11a* versus 0.10, equivalent to 5 sites, for *Pg11c*). The amino acid site detected under positive selection in *Pg11a*, with a probability of 0.98, is at position 141 and corresponds to a Glycine (G) in the precursor of the protein in *M. truncatula* [GenBank:AES65910]. At this position, *M. ciliaris*, *M. polymorpha* and *M. ruthenica* have a Lysine (K), *M. rigiduloides* a Serine (S) and *M. coerulea* an asparagine (N). Five amino acid positions under positive selection were detected in *Pg11c*. None of these 5 amino acid positions is the same than that detected in *Pg11a*. Three amino acid positions had an estimated posterior probability to be under positive selection greater than 0.95: position 110 (a Glutamic acid, E), position 132 corresponding to a Glutamine (Q) and position 303 corresponding to a Threonine (T) (position on the precursor protein in *M. truncatula*, [GenBank:AES65907]). At position 110, *M. littoralis*, *M. tricycla* and *M. tornata* have an Aspartic acid (D), and *M. polymorpha* an Asparagine (N). At position 132, the Glutamine of *M. truncatula* changes for a Threonine (T) in *M. rigiduloides* and *M. noeana* and for an Alanine (A) in *M. polymorpha* and *M. coerulea*. Finally, at position 303, all the species have a Methionine (M), except *M. truncatula* which has a Threonine (T), *M. tricycla* and *M. tornata* a Leucine (L) and *M. polymorpha*

Table 1 Branch models: estimated parameters and log-likelihood ratio tests

Paralog pair	Model	log*L*	np	Branchs	ω	LRT		*p*-value
Lax2/Lax4	M_0	−2847.12	58	OG	0.05			
				Lax	0.08			
	M_P	−2846.50	59	OG	0.05			
				Lax2	0.07	vs. M_0	1.24	0.26
				Lax4	0.10			
	M_A	−2840.62	59	OG	0.05			
				Lax$_{\text{early}}$	0.14	vs. M_0	13.0**	0.00031
				Lax$_{\text{late}}$	0.05			
	M_{PA}	−2839.60	61	OG	0.05			
				Lax2$_{\text{early}}$	0.19	vs. M_P	13.8**	0.001
				Lax4$_{\text{early}}$	0.11			
				Lax2$_{\text{late}}$	0.04	vs. M_A	2.04	0.36
				Lax4$_{\text{late}}$	0.06			
Pg3/Pg11	M_0	−4127.20	62	OG	0.06			
				Pg	0.33			
	M_P	−4125.06	63	OG	0.06			
				Pg3	0.25	vs. M_0	4.28*	0.04
				Pg11	0.41			
	M_A	−4124.78	63	OG	0.06			
				Pg$_{\text{early}}$	0.22	vs. M_0	4.84*	0.03
				Pg$_{\text{late}}$	0.38			
	M_{PA}	−4123.22	65	OG	0.06			
				Pg3$_{\text{early}}$	0.24	vs. M_P	3.68	0.16
				Pg11$_{\text{early}}$	0.21			
				Pg3$_{\text{late}}$	0.29	vs. M_A	3.12	0.21
				Pg11$_{\text{late}}$	0.44			
Pg11a/Pg11c	M_0	−3604.36	40	OG	0.27			
				Pg11	0.44			
	M_P	−3603.75	41	OG	0.27			
				Pg11a	0.38	vs. M_0	1.22	0.27
				Pg11c	0.50			
	M_A	−3604.19	41	OG	0.27			
				Pg11$_{\text{early}}$	0.50	vs. M_0	0.34	0.57
				Pg11$_{\text{late}}$	0.42			
	M_{PA}	−3603.39	43	OG	0.27			
				Pg11a$_{\text{early}}$	0.50	vs. M_P	0.72	0.70
				Pg11c$_{\text{early}}$	0.54			
				Pg11a$_{\text{late}}$	0.35	vs. M_A	1.60	0.45
				Pg11c$_{\text{late}}$	0.49			

[*] Note. Models: for M_0, ω is allowed to take a different value only in the branch of the outgroup (OG); for M_P, M_A and M_{PA} ω is allowed to take a different values according to the tested effect, *i.e.* “paralogs”, “age” or combined as explained in Figure 1. np: number of free parameters; log*L*: log-likelihood; LRT: likelihood ratio test statistic between indicated models; one (respectively two) asterisk indicates that the probability of observing such an LRT or higher under the compared model is <0.05 (respectively <0.01), assuming that the LRT follows a χ^2 distribution with the difference of free parameters between the compared models as the number of degrees of freedom.

Table 2 Site models results: estimated parameters and log-likelihood ratio tests

Gene	Model	np	log*L*	Parameters	LRT		*p*-value
Lax2	M_{8A}	35	−1600.22	p = 0.01 q = 2.86 ω_{ad} = 1 p_{ad} = 0.03			
	M_8	36	−1600.19	p = 0.01 q = 3.00 ω_{ad} = 1.10 p_{ad} = 0.03	vs. M_{8A}	0.04	0.84
Lax4	M_{8A}	23	−1391.38	p = 6.30 q = 99.00 ω_{ad} = 1 p_{ad} = 0.00			
	M_8	24	−1391.38	p = 6.30 q = 99.00 ω_{ad} = 1.00 p_{ad} = 0.00	vs. M_{8A}	0.00	1
Pg3	M_{8A}	23	−1616.24	p = 4.89 q = 99.0 ω_{ad} = 1.00 p_{ad} = 0.24			
	M_8	24	−1615.18	p = 0.13 q = 0.40 ω_{ad} = 4.88 p_{ad} = 0.01	vs. M_{8A}	2.11	0.15
Pg11	M_{8A}	39	−2221.99	p = 2.16 q = 99.00 ω_{ad} = 1 p_{ad} = 0.36			
	M_8	40	−2210.38	p = 0.01 q = 20.01; ω_{ad} = 6.30 p_{ad} = 0.04	vs. M_{8A}	23.22**	1.44 10^{-6}
Pg11a	M_{8A}	19	−1388.97	p = 2.32 q = 89.36 ω_{ad} = 1.00 p_{ad} = 0.35			
	M_8	20	−1385.65	p = 0.47 q = 0.96 ω_{ad} = 11.61 p_{ad} = 0.02	vs. M_{8A}	6.64**	9.99 10^{-3}
Pg11c	M_{8A}	21	−1519.41	p = 0.01 q = 2.54 ω_{ad} = 1.00 p_{ad} = 0.32			
	M_8	22	−1512.06	p = 0.01 q = 0.05 ω_{ad} = 4.45 p_{ad} = 0.10	vs. M_{8A}	14.69**	1.26 10^{-4}

Note. Models are M_{8A}, ω following a β distribution discretized into 10 categories of similar frequency ($0 < \omega_{1-10} < 1$) plus an additional category of $\omega_{ad} = 1$, accounting for neutral sites; M_8 differs from M_{8A} only by the additional category of ω which is constraint to be superior to 1 ($\omega_{ad} > 1$), to account for sites under positive selection.

Parameters are frequencies and values of ω for M_{8A} and M_8, p and q are the parameters in β distribution; for M_{8A} and M_8 p_{ad} and ω_{ad} are the frequencies and values of additional class of ω; NB ω_{ad} is fixed equal to one in $M_{8A.}$ np: number of free parameters; log*L*: log-likelihood; LRT: likelihood ratio test statistic between indicated models; one (respectively two) asterisk indicates that the probability of observing such an LRT or higher under the compared model is <0.05 (respectively <0.01), assuming that the LRT follows a χ^2 distribution with the difference of free parameters between the compared models as the number of degrees of freedom.

a Lysine (K). The two other sites had a posterior probability of 0.997 and 0.996, on a Tryptophan (T) in position 161 and an Alanine (A) in position 270, respectively. At position 161, *M. rigiduloides*, *M. noeana*, *M. coerulea* and *M. ruthenica* have a Histidine (H), *M. polymorpha* a (R), and *M. rigidula* an Asparagine (N). Finally, at position 270, *M. littoralis*, *M. tricycla*, *M. rigiduloides* and *M. rigidula* have a Serine (S), and *M. noeana* a Glycine (G).

Selective pressures along branches and sites of each paralog: testing for a "paralog" effect

In the third model we used, the clade model M_D [33], ω varies among sites (with either two or three categories) and selective pressure at one class of sites is allowed to differ in the two clades of the phylogeny (Figure 2c). We tested the significance of M_D models, with two (or three) categories of sites, compared to null M_3 models (discrete model), which assume that two (or three) classes of sites are evolving under different levels of selective pressures, but without difference between clades. As in the previous sections, when the *Pg11*/*Pg3* paralogs pair was analysed, both the sequences of *Pg11a* and *Pg11c* were considered for *Pg11*.

For the three paralogous gene pairs studied, models M_D for which one class of ω is allowed to differ between paralogous gene clades were significantly better than null models M_3 in which no variation between clades is allowed (Table 3). For the *Lax2* and *Lax4* paralogs tests comparing M_D and M_3 were significant when either two or three categories of ω were considered. For the other two pairs of paralogs, the test comparing M_D and M_3 was significant only when both models were defined with two categories of ω. These results revealed, for each pair, the presence of sites evolving under divergent selective pressures between the paralogous gene clades.

For *Lax2*/*Lax4*, none of the ω values was larger than 1, consistently with the result of the M_8 versus M_{8A}

comparison. The model with three categories indicates that more than 77% of amino acid positions are very strongly constrained (ω very close to 0). The other two categories are allowed to vary between the two clades. For *Lax2* a small proportion of sites is neutrally evolving (ω ~ 1) and the rest is mildly constrained (ω = 0.55), whereas in *Lax4* both categories are effectively neutral (ω = 0.97).

For *Pg11*/*Pg3* and *Pg11a*/*Pg11c*, the category of sites fixed across clades were also found to be under purifying selective pressure (ω = 0.09 and 0.12, respectively). When 2 categories of ω were considered, M_D was significantly better than the null model M_3 (LRT = 6.16, $p = 0.01$ and LRT = 4.45, $p = 0.03$ for the *Pg11*/*Pg3* and *Pg11a*/*Pg11c* paralogs pairs respectively). The category of sites allowed to differ in M_D model had a proportion of 32% and appeared to be nearly neutrally evolving in *Pg3* (ω = 0.74) but under positive selection in *Pg11* (ω = 1.35), as found with the M_8 model. Concerning the *Pg11a*/*Pg11c* paralogous gene pair and as previously detected with site models, the M_D model revealed that positive selection occurs for the *Pg11a* gene and for the *Pg11c* gene (ω = 1.54 and 3.13 respectively), but in addition M_D actually detected a difference in the rate of positive selection between the paralogous copies, which appeared to be stronger in *Pg11c*.

Discussion

In this paper we examined patterns of molecular evolution of three paralogous gene pairs, in order to detect signatures of post-duplication functional divergence. We chose a time scale that allows analysing patterns of natural selection by examining patterns of nucleotide substitution of protein-coding sequences. With that aim, we focused on three sets of paralogs from the *Medicago truncatula* genome, *Lax2*/*Lax4*, *Pg3*/*Pg11* and *Pg11a*/*Pg11c*. The duplications leading to these sets of paralogs occurred before the radiation of the 17 species studied but are still recent, as the three set of paralogs, *Lax2*/*Lax4*, *Pg3*/*Pg11* and *Pg11a*/*Pg11c,* exhibit still 83, 72 and 88% nucleotide identity, respectively. Furthermore, we selected genes that are putatively involved in symbiotic functions, considering that interspecific interactions can involve both evolution of novelty (especially in the case of the legume-rhizobium symbiosis which evolved relatively recently) and co-evolutionary phenomena that are detectable through signatures of positive selection.

Table 3 Branch-site models: estimated parameters and log-likelihood ratio tests

	Model (k)	np	Log*L*	LRT		*P*-value	prop	clade	ω
	M_3 (2)	58	−2375.94						
Lax2/Lax4	M_D (2)	59	−2369.54	vs. M_3 (2)	12.8**	$3.47\ 10^{-4}$	0.76		0.00
							0.24	*Lax2*	0.15
								Lax4	0.52
	M_3 (3)	60	−2375.94						
	M_D (3)	61	−2361.85	vs. M_3 (3)	28.17**	$1.11\ 10^{-7}$	0.77		0.005
							0.03		0.97
							0.20	*Lax2*	0.55
								Lax4	0.97
	M_3 (2)	61	−3411.03						
Pg11/Pg3	M_D (2)	62	−3407.95	vs. M_3 (2)	6.16*	0.01	0.68		0.09
							0.32	*Pg3*	0.74
								Pg11	1.35
	M_3 (3)	63	−3400.58						
	M_D (3)	64	−3399.59	vs. M_3 (3)	1.98	1.16			
	M_3 (2)	40	−2216.23						
Pg11a/Pg11c	M_D (2)	41	−2214.01	vs. M_3 (2)	4.45*	0.03	0.79		0.12
							0.21	*Pg11a*	1.54
								Pg11c	3.13
	M_3 (3)	42	−2210.38						
	M_D (3)	43	−2209.37	vs. M_3 (3)	2.01	0.16			

Note. Models: for M_3 (discrete model), ω is free to take the number of values indicated in brackets (k); these values are homogenous in all branches of the tree; for M_D, as explained in Figure 1 (b), one category of ω is allowed to differ between the two paralogous genes clades of the tree [33]. np: number of free parameters; log*L*: log-likelihood; LRT: likelihood ratio test statistic between indicated models; one (respectively two) asterisk indicates that the probability of observing such an LRT or higher under the compared model is <0.05 (respectively <0.01), assuming that the LRT follows a χ^2 distribution with the difference of free parameters between the compared models as the number of degrees of freedom.

Models describing the evolutionary fate of duplicated genes once the duplication is fixed in the species suppose different forms of selective pressures [11]. First, according to the neofunctionalization model, *i.e.* evolution of a new function through functional divergence of one of the duplicated copies, selective pressures are expected to be asymmetrical between paralogs [1]. The copy fulfilling the ancestral function is expected to remain under purifying selection while the other copy is expected to experience a short period of relaxed constraint and then positive selection driving the acquisition of its new function. Second, the subfunctionalization model envisions the fixation of complementary degenerative mutations [12]. Under this model, relaxation of purifying selection is expected during the period of functional redundancy, and may allow the fixation of at least two complementary degenerative mutations (one in each gene). When both copies are jointly required to fulfil the ancestral gene function, purifying selection is still expected to be prevalent to maintain both copies. Although both models have been functionally validated, they are not exclusive and more complex scenarios combining the steps cited previously have been devised [15,25].

For all three studied paralogous gene pairs, the two copies exhibit different regimes of selection. This result suggests that these paralogous gene pairs have undergone at least some functional differentiation. Three different tests were used to qualify selective pressures governing the paralogs. The first one contrasted the average ω between paralog clades of the phylogeny and yielded significant differences only for *Pg11*/*Pg3* (Table 1). The second test is specifically designed to detect positive selection affecting only a few sites of the sequence. We found signatures of positive selection in both *Pg11a* and *Pg11c* copies, and in *Pg11* (Table 2). Finally, the clade model (Table 3) is a combination of branch and site models and allows investigating specifically the presence of sites evolving under divergent selective pressures between the paralogous genes and quantify its proportion. The clade model (M_D) detected a significant increase of ω in *Pg11* due to the occurrence of positive selection, as detected by the site model M_8. For the paralogous pair *Pg11a/Pg11c*, branch models failed to detect any difference in selective pressure. Model D is more detailed and allows showing that sites under positive selection actually experience a stronger positive pressure in *Pg11c* than in *Pg11a*. *Lax2* is the subject of an intense purifying selection whereas *Lax4* harbours some sites (20% of sites) evolving quasi neutrally ($\omega = 0.97$). The combination of these different tests provides a more complete picture of the selective pressures at work on each set of paralog. Since each single test addresses a single hypothesis, the comparison of several complementary tests allows acquiring a more complete picture. However, the clade model, which accounts for both variation of ω among branches and amino acid position, appears as the most informative for qualifying changes of selective constraint during duplicated genes evolution [33]. The only drawback is that it does not test formally for positive selection.

We observed that the *Pg3* and *Pg11c* gene copies were pseudogenes in several species: in *M. littoralis* and *M. tricycla* for *Pg3* and in *M. tornata*, *M. rigidula* and *M. polymorpha* for *Pg11c*. Since the three genes are present and potentially functional in, at least, four other species among those studied, we can hypothesise that the mutations affecting the function of these gene copies occurred, in some phylum, after the two successive rounds of duplications leading to the presence of three copies. This observation suggests that redundancy between copies is sufficient to have allowed the loss of one copy in several species.

Functional redundancy generated by multiple copies also implies periods of relaxed selection pressures, except if duplication itself is advantageous as it is the case, for instance, for a positive dose effect of copy number [11]. Redundancy is expected to occur with a larger probability when divergence between copies is slowed as it is the case of gene conversion [34]. The phylogenetic miss positioning we observed for four genes copies (Figure 1) may be explained by gene conversion. One way to test this hypothesis would be to sequence other individuals of *M. tornata* for example, in order to see if we could detect shared polymorphism between copies, which is a signature of gene conversion [11].

We detected sites under positive selection in *Pg11* but not in *Pg3*. Rodriguez-Llorente *et al.* [26] suggested that *Pg3* has been recruited by symbiosis after a duplication affecting an ancestral pollen-specific gene. The authors suggested that the modifications occurred essentially in the promoter region. Our results show that positive selection targeted both copies of *Pg11* independently, possibly indicating the evolution of novel gene function. The polygalacturonase family contains members in organisms as distantly related as plants and eubacteria. In plants this gene family has been expanding dramatically through rounds of whole-genome duplications, segmental duplications and tandem duplications (66 and 59 copies in *Arabidopsis thaliana* and rice respectively) [35]. The high level of expansion of this family, generating periods of high redundancy, was probably accompanied by pseudogenization events, equivalent to those we detected in the *Medicago* genus. However as expression patterns are diverse between members of the family [35] subfunctionalization events were probably involved in the overall high retention rate of functional genes, notable in this family. Functional divergence among members of large gene families may also be driven by positive selection. Main examples in plants are disease resistance genes [36], transcription factors [37] or genes involved in development

[38]. In our study, positive selection is detected in *Pg11*, resulting from the cumulative effects of positive selection in both *Pg11a* and *Pg11c*, the more recent duplicated gene pair we studied. Actually, this mode of selection does correspond to neither neofunctionalization nor subfunctionalization in their stricter definition. Subfunctionalization does not predict positive selection in either copy, while neofunctionalization predicts positive selection in only one copy (if detectable). Both copies could be under positive selection because they inherited, from the ancestral *Pg11* gene, functions that imply regime of positive selection. Alternatively, neo-functionalization could involve adaptive differentiation of both copies (to avoid functional overlap), that would mediate adaptive evolution of both copies. Selection targets different sites in *Pg11a* and *Pg11c* and the strength of positive selection is different between them (Table 3). This observation is compatible with both models.

According to the clade models, the paralogs *Lax2* and *Lax4* experience different modes of selection. Both genes are mainly under purifying selection. Interestingly no pseudogenes were detected in *Lax2* or in *Lax4*. The redundancy stage subsequent to the duplication generating *Lax2* and *Lax4* is not detectable anymore and may have been shorter than in *Pg* gene family. However, *Lax4* appeared to be slightly, but significantly, less constrained than *Lax2*. According to the clade models (with 2 or 3 classes of sites, Table 3) a relaxation of constraint is observed for about 20% of the sites for *Lax4* relative to *Lax2*. This means either that *Lax4* acquired a function that implies less functional constraints or that both genes underwent subfunctionalization in such a way that the protein sequence of *Lax4* is less constrained. Currently, the precise functions of *Lax2* and *Lax4* are not known. Both paralogs are expressed in shoot and roots of nodulating plants of *M. truncatula*. *Lax2* is found in Expressed Sequence Tag (EST) libraries built from different tissues (2 in early seed development, 2 in flowers, early seeds, late seeds and stems, 2 in mixed root and nodules, 1 in nematode-infected roots, in developing flowers and phosphate-starved leaf). *Lax4* is not found in EST libraries but expression of *Lax4* was detected in shoots and roots of nodulating plants of *M. truncatula* [39].

The models contrasting ω in different branches allowed testing transient relaxation of purifying selection predicted to occur immediately after duplication. A significant increase of ω was detected in basal branches of the *Lax2*/*Lax4* phylogeny. The opposite trend was detected for the *Pg11*/*Pg3* pair, where purifying selection appeared to be actually weaker in late branches than in early branches, particularly for *Pg11* (ω = 0.44). However, the value of ω in late branches was likely biased by the occurrence of positive selection in *Pg11*, because branch models average over all sites.

Conclusions

This study illustrates the multiplicity of mechanisms governing the evolutionary fate of duplicated genes and, in particular, the relative age of the duplication. Analysis of nucleotide substitution rates in gene coding sequence can discriminate between qualitative phenomenon (occurrence of positive selection) or quantitative differences (levels of ω between clades and its variation among branch and sites). Further studies of the factors governing evolution of duplicated genes will benefit from taking into account features of the evolution of gene families involving successive rounds of duplications.

Methods

Plant material

One accession was selected in sixteen diploid species of the *Medicago* genus: *M. arabica, M. ciliaris, M. carstiensis, M. coerulea, M. laciniata, M. littoralis, M. marina, M. noëana, M. orbicularis, M. polymorpha, M. rigidula, M. rigiduloides, M. ruthenica, M. sauvagei, M. tornata, M. tricycla*. Accession numbers, geographic location and mating systems are presented in Additional file 2.

Selection of duplicated genes

Genes were chosen on the basis of the *Medicago truncatula* line A17 whole genome sequence [28]. We selected two multigenic families meeting exhibiting recent rounds of duplications and involved in symbiosis-related functions. First, polygalacturonases (PG) form a gene family that is ubiquitous in the plant kingdom. These proteins are involved in the degradation of polysaccharides found in higher plants cell walls. The gene *Pg11* is involved in pollen tube elongation in *M. truncatula* and is located on chromosome 2. *Pg3*, located on chromosome 5 in *M. truncatula*, has been shown to be involved in the tip growth of the infection thread during the establishment of the symbiosis with nitrogen-fixing bacteria *Sinorhizobium* sp. [26]. We also identified a more recent tandem duplication of *Pg11*, resulting in the paralogs *Pg11a* and *Pg11c*. The pairs *Pg3-Pg11c* and *Pg11a-Pg11c* exhibit respectively 72 and 88% nucleotide sequence identity, and 62 and 81% amino acid sequence identity.

Second, we chose family of auxin efflux carrier, *Lax* (Like-*Aux1*), for which five members have been identified in *Medicago truncatula* [39]. Auxin is generally involved in the control of tissue growth and in particular during the development of nodules, the symbiotic organ hosting *Sinorhizobium* symbionts [40]. Auxin is synthesized in aerial organs (leaves and shoot apex) and is directionally transported. As a result, auxin carriers such as LAX proteins play an important role in auxin-controlled processes. We chose to study the youngest paralogous gene pair *Lax2-Lax4* that presents 83% of nucleotide identity, and 87% of amino acid identity. The sequence accession numbers are AY115843 and AY115844 respectively for *Lax2* and *Lax4*.

Sequencing

To amplify specifically the coding region of each paralogous gene, we defined specific and non-specific primers. Non-specific primers were defined using the common sequence of both paralogous gene for each pair (*i.e.* not allowing to amplify separately each paralogs), whereas copy-specific primers were defined using polymorphism between the paralogs, available in the reference *Medicago truncatula* genotype A17. In a first step, specific primers combinations were used to amplify specifically each paralogs. Then, sequencing was performed using specific and/or non-specific primers. The primer sequences and their position on the genomic sequences of the five genes are available in Additional file 3 and Additional file 4 respectively. As divergence between species was often the cause of unsuccessful amplifications, several copy-specific primer pairs were defined to increase the chances of amplification in the sixteen studied species. Additional amplification rounds were performed to close sequencing gaps. For the most recent paralogs pair (*Pg11a-Pg11c*), the sequences obtained were labelled according to the primer combinations used for amplification and sequencing: when using primers designed for *Pg11a* (respectively *Pg11c*) , the sequencing product was qualified as '*Pg11a*' copy (respectively '*Pg11c*' copy).

Most sequences were obtained from genomic DNA, except for *Lax2*, which was sequenced from cDNA due to its large size. DNA extraction and genomic DNA amplifications and sequencing were performed as described in [41]. Total RNA was extracted from fresh leaves with a TRI REAGENT (T9424, Sigma®) buffer. Reverse transcription was done using the Reverse Transcription System kit from Promega®. Amplification and sequencing from cDNA were then performed as for genomic DNA. Chromatograph assembly and alignment were performed using programs of the Staden package v1.5 [42]. Visual inspection and correction of base calling and alignment were performed at this stage. The sequence editor Artemis v9 [43] was used to validate the reading frame and detect eventual frame shifts and/or premature stop codon mutation.

Outgroups

The *Medicago truncatula* sequence of gene *Lax1* [GenBank:AY115841] was used as outgroup to root the *Lax2-Lax4* pair phylogeny. *Lax1* diverged from *Lax2* and *Lax4* through a more ancient duplication [39]. Following the phylogenetic tree of the plants PG and endoglucanases published by Rodriguez-Llorente [26], we selected a PG coding sequence from *Salix gilgiana* [GenBank:AB029458] as outgroup for the Pg3-*Pg11a*/*c* phylogenetic tree. The *M. truncatula* copy of *Pg3* was used as outgroup for the *Pg11a- Pg11c* pair phylogeny.

Phylogenetic analysis

Maximum-likelihood phylogenetic trees were inferred using the PHYML program [44]. Maximum-likelihood analyses were conducted under the GTR molecular substitution model. Site to site variation in substitution rate was modeled by estimating the proportion of invariant sites and assuming that rates among the remaining sites were gamma distributed (4 categories were used to discretize the gamma distribution). The confidence level of each node was estimated using 100 bootstrap repetitions. Nucleotide and amino acid alignments of *Lax* genes are available in Additional file 5 and Additional file 6 respectively. Nucleotide and amino acid alignments of *Pg* genes are available in Additional file 7 and Additional file 8 respectively.

Variation in substitution rates was analyzed using codon substitution models where the parameter ω is defined as the ratio of non-synonymous (dN) to synonymous (dS) substitution rates [45]. We used eight models that make different assumption regarding variation of ω (Figure 2) in the phylogeny of each pair of paralogs. The first four models account of variation of ω among branches of the phylogeny [46] (Figure 2a). Model M_0 assumes a single ω value for both paralogs. Model M_P allows a "paralog effect" by assigning a different ω for each paralog clade in the tree. Model M_A allows an "age effect" and assigns a single ω for the basal (ancestral) branch of both paralogs clades and a different ω to all other (more recent) branches within both clades. Model M_{PA} allows for both levels of variation. All four models above are also specifying a specific ω parameter value on the branch leading to the outgroup of each paralog phylogeny. Total numbers of ω parameters are 2 for M_0, 3 for M_A and M_P, and 5 for M_{PA}.

Next, two models allowing for variation of ω among sites, but not among branches of the phylogenetic tree, and that are designed specifically to detect positive selection were used (Figure 2b) [47]. M_{8A} assumes that a fraction of the sites experience purifying selection of varying intensity by assuming that ω omega values follow a beta distribution ($0 < \omega < 1$). The remaining fraction of the site are assumed to evolve neutrally ($\omega = 1$). M_{8A} was used as null model for detecting positive selection by comparing its fit with M_8 in which the additional category of ω is free to take any value above 1 (positive selection). The comparison of M_8 versus M_{8A} provides a (likelihood ratio) test for the occurrence of positive selection (identified when at least some sites exhibit a $\omega > 1$). These two models were fitted separately to each paralog in the tree and excluding the outgroup, in order to detect positive selection occurring specifically on each copy.

The last two models are so called branch-site models that are combining variation of omega both among amino acid positions of the alignment and between different clades of the phylogeny, in our case each paralog clade (Figure 2c). These models allow testing variation of selective constraints between paralogous copies. Outgroup sequences are not considered in this analysis. The null model M_3 allows either two or three rate categories that are homogeneous along the tree. Model M_D (model D in [33]) allows selective pressure at one class of sites to differ in different clades of the phylogeny. Applied to our case, it is allowed to differ in each clade of paralogous gene (Figure 2c).

Maximum likelihood estimation of all model parameters was performed using the codeml software of the PAML package [48]. The different pairs of models are nested and were compared using likelihood ratio tests (LRTs).

Additional files

Additional file 1: Sequencing results. Table in PDF format presenting sequencing results for the five genes on the 17 species and GenBank accession numbers. Lengths are indicated in base pairs. The percentage that each sequence represents relative to the complete alignment is indicated in brackets when less than 100%. "*na*" and "*ns*" are indicated when an amplification failed and when the sequence was too short to be included in the analyses, respectively. Four sequences presented either point mutations resulting in a stop codon (*Pg11c* of *M. laciniata*), or a deletions inducing a frame shift in the coding sequence (*Pg11c* of *M. ciliaris*) or resulting in the appearance of a premature stop codon (for three sequences: *Pg11c* of *M. orbicularis* and *Pg3* of *M. littoralis* and *M. tricycla*) are indicated by "*pseudo*". Sequences with an unexpected position in the phylogeny are noted as "*phylo_excluded*".

Additional file 2: List of species used. Table in PDF format with list of sample used, germplasm accession number, life history, geographical area and ploidy level.

Additional file 3: List of primers used. Table in PDF format with names and sequences of primers used for amplification and sequencing.

Additional file 4: Schematic representation of genes and primers positions. Figure in PDF format with schematic representation of the intron/exon structure of the 5 sequenced genes on *M. truncatula* (A17) and position of the primers used for the amplification and sequencing, names and sequences of primers used for amplification and sequencing.

Additional file 5: Lax gene Nucleotide alignment. Nucleotide alignment of *Lax* genes in phyml format.

Additional file 6: *Lax* gene amino acid alignment. Amino acid alignment of *Lax* genes in phyml format.

Additional file 7: *Pg* gene nucleotide alignment. Nucleotide alignment of *Pg* genes in phyml format.

Additional file 8: Pg gene amino acid alignment. Amino acid alignment of *Pg* genes in phyml format.

Abbreviations

Pg: Polygalacturonase; Lax: Like-*Aux1* (auxin efflux carrier); LRTs: Likelihood ratio tests; OG: Outgroup; EST: Expressed sequence tag.

Competing interests

The authors do not have any kind of financial or non-financial competing interest to declare in relation to this manuscript.

Authors' contributions

NC and JR conceived and designed research, JHH, IH and AW acquired data, JHH and NC processed data, JHH, NC and SDM analyzed data and NC, JHH, JR, SDM and TB wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

Sequencing was funded by an "AIP Genomes" grant by INRA. This work was also supported by ARCAD, a flagship project of the Agropolis Foundation.

Author details

[1]INRA - Institut National de la Recherche Agronomique, UMR AGAP, Montpellier 34060, France. [2]INRA - Institut National de la Recherche Agronomique, UMR IAM, Nancy, France. [3]Bioinformatics Research Center (BiRC), Aarhus University, Aarhus, Denmark.

Received: 17 February 2012 Accepted: 20 September 2012
Published: 1 October 2012

References

1. Ohno S: *Evolution by gene duplication*. London: George Allen and Unwin; 1970.
2. Lynch M: *Genomic expansion by gene duplication*, The origins of genome architecture. Sunderland, Massachusetts: Sinauer Associates; 2007.
3. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. *Science* 2000, **290:**1151–1155.
4. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: Extensive duplication and reshuffing in the Arabidopsis genome. *Plant Cell* 2000, **12:**1093–1101.
5. Blanc G, Wolfe KH: Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 2004, **16:**1679–1691.
6. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, *et al*: Widespread genome duplications throughout the history of flowering plants. *Genome Res* 2006, **16:**738–749.
7. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, *et al*: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, **449:**463–467.
8. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, *et al*: Genome sequence of the palaeopolyploid soybean. *Nature* 2010, **463:**178–183.
9. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al*: The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 2006, **313:**1596–1604.
10. Wendel JF: Genome evolution in polyploids. *Plant Mol Biol* 2000, **42:**225–249.
11. Innan H, Kondrashov F: The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* 2010, **11:**97–108.
12. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 1999, **151:**1531–1545.
13. Moore RC, Purugganan MD: The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* 2005, **8:**122–128.
14. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: Role of duplicate genes in genetic robustness against null mutations. *Nature* 2003, **421:**63–66.
15. He X, Zhang J: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 2005, **169:**1157–1164.
16. Nadeau JH, Sankoff D: Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 1997, **147:**1259–1266.
17. Li WH, Yang J, Gu X: Expression divergence between duplicate genes. *Trends Genet* 2005, **21:**602–607.
18. Adams KL: Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. *J Hered* 2007, **98:**136–141.
19. Chain FJ, Ilieva D, Evans BJ: Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol* 2008, **8:**43.
20. Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, Wendel JF: Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). *Genetics* 2009, **182:**503–517.
21. Des Marais DL, Rausher MD: Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 2008, **454:**762–765.
22. Hittinger CT, Carroll SB: Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 2007, **449:**677–681.
23. Gayral P, Caminade P, Boursot P, Galtier N: The evolutionary fate of recently duplicated retrogenes in mice. *J Evol Biol* 2007, **20:**617–626.
24. Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y: Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome Biol* 2006, **7:**R13.
25. Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, de Pamphilis CW: Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol Biol Evol* 2006, **23:**469–478.

26. Rodriguez-Llorente ID, Perez-Hormaeche J, El Mounadi K, Dary M, Caviedes MA, Cosson V, Kondorosi A, Ratet P, Palomares AJ: **From pollen tubes to infection threads: recruitment of Medicago floral pectic genes for symbiosis.** *Plant J* 2004, **39:**587–598.
27. Sprent JI: **Evolving ideas of legume evolution and diversity: a taxonomic perspective on the occurrence of nodulation.** *New Phytol* 2007, **174:**11–25.
28. Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, *et al*: **The Medicago genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480:**520–524.
29. Lavin M, Herendeen PS, Wojciechowski MF: **Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary.** *Syst Biol* 2005, **54:**575–594.
30. Bena G, Jubier MF, Olivieri II, Lejeune B: **Ribosomal External and Internal Transcribed Spacers: Combined Use in the Phylogenetic Analysis of Medicago (Leguminosae).** *J Mol Evol* 1998, **46:**299–306.
31. Steele KP, Ickert-Bond SM, Zarre S, Wojciechowski MF: **Phylogeny and character evolution in Medicago (Leguminosae): Evidence from analyses of plastid trnK/matK and nuclear GA3ox1 sequences.** *Am J Bot* 2010, **97:**1142–1155.
32. Maureira-Butler IJ, Pfeil BE, Muangprom A, Osborn TC, Doyle JJ: **The reticulate history of Medicago (Fabaceae).** *Syst Biol* 2008, **57:**466–482.
33. Bielawski JP, Yang Z: **A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution.** *J Mol Evol* 2004, **59:**121–132.
34. Teshima KM, Innan H: **The effect of gene conversion on the divergence between duplicated genes.** *Genetics* 2004, **166:**1553–1560.
35. Kim J, Shiu SH, Thoma S, Li WH, Patterson SE: **Patterns of expansion and expression divergence in the plant polygalacturonase gene family.** *Genome Biol* 2006, **7:**R87.
36. Sun X, Cao Y, Wang S: **Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice.** *Plant Physiol* 2006, **140:**998–1008.
37. Jia L, Clegg MT, Jiang T: **Excess non-synonymous substitutions suggest that positive selection episodes occurred during the evolution of DNA-binding domains in the Arabidopsis R2R3-MYB gene family.** *Plant Mol Biol* 2003, **52:**627–642.
38. Yang Z, Gu S, Wang X, Li W, Tang Z, Xu C: **Molecular evolution of the CPP-like gene family in plants: insights from comparative genomics of Arabidopsis and rice.** *J Mol Evol* 2008, **67:**266–277.
39. Schnabel EL, Frugoli J: **The PIN and LAX families of auxin transport genes in Medicago truncatula.** *Mol Genet Genomics* 2004, **272:**420–432.
40. Desbrosses GJ, Stougaard J: **Root nodulation: a paradigm for how plant-microbe symbiosis influences host developmental pathways.** *Cell Host Microbe* 2011, **10:**348–358.
41. De Mita S, Santoni S, Hochu I, Ronfort J, Bataillon T: **Molecular evolution and positive selection of the symbiotic gene NORK in Medicago truncatula.** *J Mol Evol* 2006, **62:**234–244.
42. Staden R: **The Staden sequence analysis package.** *Mol Biotechnol* 1996, **5:**233–241.
43. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944–945.
44. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696–704.
45. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11:**725–736.
46. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.** *Mol Biol Evol* 1998, **15:**568–573.
47. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155:**431–449.
48. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555–556.

doi:10.1186/1471-2148-12-195
Cite this article as: Ho-Huu *et al.*: Contrasted patterns of selective pressure in three recent paralogous gene pairs in the *Medicago* genus (L.). *BMC Evolutionary Biology* 2012 **12**:195.