Reconstruire 100 millions d'années d'évolution des génomes de graminées par polyploïdisation

Arnaud Bellec

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 – Paul Sabatier

Présentée et soutenue par

Arnaud BELLEC

Le 15 décembre 2022

Reconstruire 100 millions d'années d'évolution des génomes de graminées par polyploïdisation

École doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries**

Spécialité : **Développement des plantes, interactions biotiques et abiotiques**

Unité de recherche
INRAE-CNRGV - Centre National de Ressources Génomiques Végétales

Thèse dirigée par
Nicolas LANGLADE et Jérôme SALSE

Jury
M. Christophe DUNAND, Président du jury
M. Stéphane MAURY, Rapporteur
M. Olivier PANAUD, Rapporteur
Mme Dominique THIS, Rapportrice
M. Christophe PLOMION, Examinateur
Mme Maria MANZANARES-DAULEUX, Examinatrice
M. Nicolas LANGLADE, Co-directeur de thèse
M. Jerome SALSE, Co-directeur de thèse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITE DE TOULOUSE

Délivré par l'Université Toulouse 3 – Paul Sabatier

Présentée et soutenue par

Arnaud BELLEC

Le 15 décembre 2022

Reconstruire 100 millions d'années d'évolution des génomes de graminées par polyploïdisation

École doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries**

Spécialité : **Développement des plantes, interactions biotiques et abiotiques**

Unité de recherche

INRAE-CNRGV - Centre National de Ressources Génomiques Végétales

Thèse dirigée par

Nicolas LANGLADE et Jérôme SALSE

Jury

M. Christophe DUNAND, Président du jury
M. Stéphane MAURY, Rapporteur
M. Olivier PANAUD, Rapporteur
Mme Dominique THIS, Rapportrice
M. Christophe PLOMION, Examinateur
Mme Maria MANZANARES-DAULEUX, Examinatrice
M. Nicolas LANGLADE, Co-directeur de thèse
M. Jerome SALSE, Co-directeur de thèse

Remerciements

Je remercie l’ensemble des membres du jury qui ont évalué mon travail. Merci à Dominique This, Stéphane Maury et Olivier Panaud qui ont accepté d'être les rapporteurs. Merci à Maria Manzanares-Dauleux, Christophe Plomion et Christophe Dunand d'avoir fait partie du jury.

Un remerciement, bien évident spécial, à Hélène Bergès, avec qui j'ai travaillé pendant de nombreuses années, qui m'a encouragé à entamer ce long travail de thèse.

Un remerciement particulier à Jérôme Salse, codirecteur de thèse, pour m'avoir accueilli au sein de l'équipe Paléo-EVO, pour ses connaissances et sa disponibilité constante en dépit des responsabilités qui sont les siennes. Essentiel tout au long de ce travail, il m'a montré la voie.

Merci à Nicolas Langlade, codirecteur de thèse, pour avoir été là pour me permettre de poursuivre cette thèse et pour son regard toujours bienveillant.

Je remercie l’école doctorale SEVAB qui m’a accompagné en prenant en compte mon statut particulier.

Je remercie vivement l'équipe Paléo-EVO. Notamment, Mamadou Dia-Saw et Wandrille Duchemin pour les nombreuses interactions et les discussions très riches que nous avons eues. Caroline Pont et Cécile Huneau m’ont dispensé leur savoir sur la reconstruction des génomes et m’ont accueilli chaleureusement lors de mes séjours au GDEC. David Armisen, Emile Mardoc et Peter Civan ont apporté leurs compétences et leurs regards critiques, particulièrement pour l'écriture de la publication.

Au CNRGV, je remercie tout d'abord, Sonia Vautrin, Laetitia Lançon et Stéphane Cauet, mes acolytes de la direction collégiale, qui ont fait preuve de patience et de soutien alors que les tâches qui nous incombaient étaient nombreuses. Merci à Margaux-Allison Fustier pour son aide pour l'analyse des données. Merci aux collègues de l'équipe qui par leur écoute et leur curiosité, m'ont aidé à avancer : Nathalie Rodde, Caroline Callot, Isabelle Dufau, Nadine Gautier, Elisa Prat, Nadège Arnal, Joëlle Fourment, William Marande et David Pujol, sans oublier les nouveaux arrivants, Charlotte Cravero, Anthony Théron et Robin Schwartz.

Je remercie également ma famille et mes amis qui, lorsque nous nous retrouvions, ne manquaient jamais de me demander des nouvelles de mon mystérieux travail de thèse.

Enfin, je remercie et je dédie cette thèse à mes parents Cécile et Yves, à ma compagne Erika et à mes enfants Erin et Colin.

Résumé

La polyploïdie, résultat de la duplication du génome entier d'un être vivant (*whole genome duplication*, WGD), est un phénomène omniprésent chez les plantes. À la suite de sa découverte, il y a plus d'un siècle, la polyploïdie a fait l'objet de nombreuses études qui ont montré qu'elle constitue une force majeure permettant aux plantes de s'adapter à leur environnement. La compréhension des mécanismes qui expliquent son apparition et ses conséquences ont progressé de façon spectaculaire avec l'avènement du séquençage des plantes. Disposer des séquences de multiples génomes de plantes a permis de retracer l'histoire évolutive des plantes grâce aux analyses de paléogénomique et de mettre en évidence des évènements de polyploïdisation anciens. L'observation de l'évolution des caryotypes a révélé l'existence de mécanismes remodelant la structure des chromosomes (inversion, fusion, fission). La comparaison des contenus en gènes des espèces modernes avec ceux de leurs ancêtres a mis en évidence des phénomènes de pertes de gènes spécifiques suivant un évènement de polyploïdisation. Les données omiques (expression, méthylation) ont apporté de nouveaux éléments pour comprendre les conséquences de polyploïdie sur la régulation et la physiologie. Cependant, selon les méthodes d'analyse utilisées, en fonction des espèces végétales considérées et du type de données considérées, les observations des conséquences de la polyploïdie et les théories sur les règles qui régissent l'évolution des génomes polyploïdes divergent selon les auteurs. Les conclusions visant à identifier des patrons génériques susceptibles de régir l'évolution des génomes de plantes post-polyploïdie demeurent largement spéculatives.

Cette étude propose une analyse des mécanismes actionnés par la polyploïdie en considérant un panel de 8 espèces de *Poaceae*, une famille du groupe des monocotylédones comptant des espèces d'importance économique majeure comme le blé, le riz, l'orge et le maïs, et présentant plusieurs évènements de polyploïdisation depuis 100 millions d'années. L'analyse comparative des données omiques disponibles pour ces espèces lève le voile sur les bases structurales (inversion, fusion, fissions, duplications), le destin des gènes dupliqués et la régulation (expression et méthylation) des génomes. Ces résultats permettent de proposer un scénario évolutif post-polyploïdie pour les céréales, identifiant les forces et les évènements majeurs, ainsi que leur séquence temporelle.

L'étude démontre que les lignées polyploïdes issues d'un événement de WGD commun présentent souvent les mêmes schémas de changements structuraux et de dynamique évolutive. Ces patrons sont cependant difficiles à généraliser dès lors que des événements de WGD indépendants sont considérés, principalement à cause des impacts d'autres forces propres à l'histoire évolutive des espèces telles que la sélection et la domestication des cultures. Si la polyploïdie est liée sans équivoque au succès évolutif des graminées depuis 100 millions d'années, il demeure illusoire d'attribuer ce succès à des effets récurrents de la polyploïdisation. Il apparait plutôt que la polyploïdie augmentant le potentiel adaptatif des néopolyploïdes offre des possibilités diverses d'adaptation au changement de l'environnement qui se réalisent de manière spécifique chez les différentes espèces végétales. Globalement, cette étude démontre clairement que la reprogrammation génomique post-polyploïdie est plus complexe que ce qui a été traditionnellement rapporté dans l'étude d'une espèce. Elle ouvre la voie à une comparaison critique et complète entre espèces polyploïdes de branches évolutives indépendantes.

Summary

Polyploidy, the result of a whole genome duplication (WGD) event, constitutes a ubiquitous phenomenon in plants. Following its discovery more than a century ago, polyploidy has been widely studied, establishing WGD as a major force allowing plants to adapt to their environment. The understanding of the mechanisms that explain its occurrence and its consequences has progressed dramatically with the advent of plant sequencing. The availability of multiple plant genome sequences has allowed us to trace the evolutionary history of plants through paleogenomic analyses and to highlight ancient polyploidization events. The analysis of the evolution of karyotypes has revealed the existence of various mechanisms that reshape of the structure of chromosomes (inversion, fusion, fission). The comparison of the gene contents of modern species with those of their ancestors has revealed the loss of specific genes following a polyploidization event. Omics data (expression, methylation) have provided new insights to understand the consequences of polyploidy on regulation and physiology. However, the consequences of polyploidy and the theories on the rules that govern the evolution of polyploid genomes differ according to the authors, depending on analysis methods used, plant species data considered. The propositions of generic patterns that may govern the evolution of post-polyploidy plant genomes remain largely speculative.

This study proposes an analysis of the mechanisms driven by polyploidy by considering a panel of 8 species of Poaceae, a family of the monocotyledonous group including species of major economic importance such as wheat, rice, barley and maize, and presenting several polyploidization events since 100 million years. The comparative analysis of the available omics data for these species reveals the structural basis (inversion, fusion, fissions, duplications), the fate of duplicated genes and the regulation (expression and methylation) of genomes. These results allow us to propose a post-polyploidy evolutionary scenario for cereals, identifying the major forces and events, as well as their temporal sequence.

The study demonstrates that polyploid lineages derived from a common WGD event often exhibit the same patterns of structural change and evolutionary dynamics. However, these patterns are difficult to generalize when independent WGD events are considered, mainly because of the impacts of other forces in the evolutionary history of the species such as selection and crop

domestication. While polyploidy is unequivocally linked to the evolutionary success of grasses over the past 100 million years, it is illusive to attribute this success to recurrent effects of polyploidization. Rather, it appears that polyploidy increases the adaptive potential of neopolyploids, which provides diverse opportunities for adaptation to environmental change that are realized in specific ways in different plant species. Overall, this study clearly demonstrates that post-polyploidy genomic reprogramming is more complex than traditionally reported in the study of one species. It paves the way for a comprehensive and critical comparison of polyploid species of independent evolutionary branches.

Contexte de la thèse

Je travaille depuis 20 ans dans le domaine l'étude de l'ADN des plantes grâce aux outils de la biologie moléculaire puis de la génomique. Après un DESS à l'université de Paris 7 j'ai d'abord travaillé à l'INRA d'Évry de 2001 à 2004. Au sein de l'URGV, unité pionnière en génomique végétale, j'ai participé dans l'équipe de Boulos Chalhoub, à la production de plusieurs banques d'ADN de blé pour isoler des régions génomiques d'intérêts. À l'époque, isoler, séquencer et assembler une région de 150 kb contenant quelques gènes chez le blé, constituait un défi en soi. En 2004, j'ai rejoint l'INRA de Toulouse et le Centre National de Ressources Génomiques Végétales, tout juste créé par Hélène Bergès. J'ai obtenu un poste permanent d'ingénieur d'étude et contribué à développer cette unité de service ayant pour mission de centraliser, conserver, diffuser et valoriser les banques d'ADN de plantes, dont certaines avaient justement été construites à l'URGV. Je me suis épanoui dans les fonctions d'ingénieur, tantôt à mettre en œuvre des protocoles de biologie moléculaire, tantôt à programmer des robots pour traiter les échantillons, toujours au cœur de l'équipe du CNRGV et au service de la communauté scientifique. À cette époque, produire un travail de thèse constituait à mes yeux un Everest hors d'atteinte. Cependant, à force d'interagir avec les chercheurs qui collaboraient avec le CNRGV, l'idée est apparue et, quelques années plus tard, le projet a débuté sous la direction d'Hélène Bergès et de Jérôme Salse, responsable de l'équipe Paléo-EVO dans l'unité GDEC à Clermont-Ferrand. Il s'agissait, en parallèle de mes activités au laboratoire, d'étudier l'évolution de régions homéologues au sein du blé hexaploïde. Des objectifs qu'ensemble, nous réviserions ensuite, en fonction des progrès technologiques et des avancées du séquençage des génomes végétaux. En dépit de mon intérêt pour le sujet, passer de mes certitudes d'ingénieur à un état d'esprit propice à la recherche s'est révélé compliqué. Là où d'ordinaire, je savais quel résultat attendre et j'avais confiance dans le fait qu'avec de la méthode et de l'assiduité, il serait possible de l'obtenir, je me suis retrouvé face à des résultats inattendus, parfois en contradiction avec mes idées initiales. C'est grâce aux échanges avec Jérôme Salse et les collègues de son équipe que j'ai mesuré la valeur des résultats obtenus et, au-delà, commencé à comprendre les ressorts de la recherche scientifique. Outre la présentation des données et des connaissances produites, c'est aussi ce parcours personnel qui transparaît dans ce manuscrit.

Abréviations

AGK : *Ancestral Grass Karyotype*

ATK : *Ancestral Triticeae Karyotype*

BAC : *Bacterial Artificial Chromosome*

BLAST : *Basic Local Alignment Score Tool*

CALP : *Cumulative Alignment Length Percentage*

CIP : *Cumulative Identity Percentage*

CNV : *Copy Number Variation*

COS : *Conserved Orthologous Set*

DD : *Degree Day*

ELD : *Expression Level Dominance*

ET : élément transposable

GWAS : *Genome-Wide Association Study*

HGP : *Human Genome Project*

IRGSP : *International Rice Genome Sequencing Project*

IWGSC : *International Wheat Genome Sequencing Consortium*

LF : *Less Fractionated*

MF : *More Fractionated*

MTP : *Minimal Tilling Path*

mya : *million years ago*

NGS : *Next Generation Sequencing*

PAV : *Presence/Absence Variation*

RNA-seq : *RNA sequencing*

RPKM : *Reads Per Kilobase per Million*

RPM : *Reads Per Million*

SNP : *Single Nucleotide Polymorphism*

TPM : *Transcripts Per Million*

WGD : *Whole Genome Duplication*

WGT : *Whole Genome Triplication*

Table des matières

I. INTRODUCTION

INTRODUCTION

1 Génomique des angiospermes

1.1 Les angiospermes : histoire évolutive et classification

1.1.1 Emergence des angiospermes au sein des plantes terrestres

La conquête du milieu terrestre par les plantes débute au paléozoïque, il y a environ 500 millions d'années, lors de l'émergence des embryophytes au sein du groupe des *Viridiplantae* dont font également partie les algues vertes. Les embryophytes, ou plantes terrestres, comprennent les groupes des bryophytes, des lycophytes, des monilophytes et des spermaphytes, ou plantes à graines, lui-même subdivisé en gymnospermes et angiospermes, ou plantes à fleurs (Figure 1).

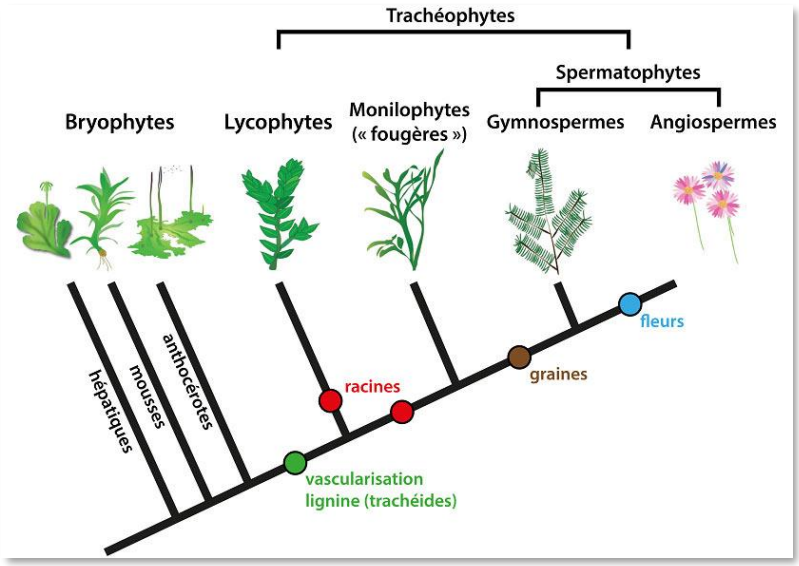


Figure 1 : arbre évolutif des embryophytes d'après P. Guériau et J. Mondejar Fernandez 2019, Les premiers écosystèmes terrestres https://www.encyclopedie-environnement.org/vivant/premiers-ecosystemes-terrestres/

Ce dernier groupe compte environ 350 000 espèces et représente plus de 90% des plantes terrestres dont dépend directement ou indirectement la majeure partie de la vie sur les terres émergées (Sauquet *et al.*, 2017) notamment grâce à la capacité de ces espèces à capter le CO_2 atmosphérique. Les angiospermes sont des plantes vasculaires dont l'ovule fécondé se développe en une graine dans un ovaire creux et fermé, généralement enfermé dans une fleur. Celle-ci porte les organes reproducteurs mâles ou femelles, ou les deux à la fois. Les fruits résultent de la transformation des organes floraux après fécondation et sont caractéristiques des angiospermes.

Au contraire, chez les gymnospermes (conifères, cycas et ginkgos), l'autre groupe majeur de plantes à graines, celles-ci ne se développent pas enfermées dans un ovaire mais sont exposées sur la surface des structures de reproduction, telles que les cônes.

A la différence des plantes non vasculaires, chez lesquelles toutes les cellules du corps végétal participent à toutes les fonctions nécessaires pour soutenir, nourrir et assurer la croissance de la plante (notamment la nutrition, la photosynthèse et la division cellulaire), les angiospermes présentent des cellules, des tissus et des organes spécialisés pour remplir chacune de ces fonctions. Cette spécialisation morphologique s'organise autour des tissus vasculaires (xylème et phloème) qui transfèrent l'eau et les nutriments dans toutes les zones de la plante. Elle définit trois types d'organes caractéristiques, à savoir, un système racinaire étendu qui ancre la plante et absorbe l'eau et les minéraux du sol, une tige qui soutient la plante et un système foliaire qui constitue le principal site de la photosynthèse pour la plupart des angiospermes. A partir de cette organisation commune, les angiospermes se sont diversifiées en une multitude de formes remarquables. A ce titre, la gamme des tailles, de la lentille d'eau, *Wolffia arrhiza,* mesurant 2 millimètres à *Eucalyptus regnans*, atteignant 100 mètres de hauteur, illustre la multiplicité des trajectoires évolutives des différentes espèces d'angiospermes. Cette diversité permet aux angiospermes d'être présentes sur l'ensemble des terres émergées de la planète, à l'exception de l'Antarctique continental, ainsi que dans les milieux aquatiques sous forme de plantes flottantes (*Nymphaea*) ou submergées (*Zosteraceae*). Du fait de leur disponibilité et de leur diversité, les angiospermes constituent une source basale de nourriture directe ou indirecte pour la majorité des êtres vivants hétérotrophes dont les animaux, humains inclus (Purugganan and Jackson, 2021). Elles sont également le groupe de plantes le plus important sur le plan économique, notamment pour la production de bois, la pharmacologie et le marché des plantes ornementales.

1.1.2 Classification phylogénique des angiospermes

Les angiospermes sont divisées en deux groupes sur la base du nombre de cotylédon au stade embryonnaire, définissant le groupe des monocotylédones et celui des dicotylédones (Figure 2). Le groupe des monocotylédones comprend les graminées, palmiers, bananiers, joncs et orchidées qui présentent un seul cotylédon, tandis que le groupe des dicotylédones regroupe

les *Asteridae*, *Caryophyllidae*, *Dilleniidae*, *Hamamelidae*, *Rosidaea* et *Magnoliidae* qui présentent deux cotylédons.

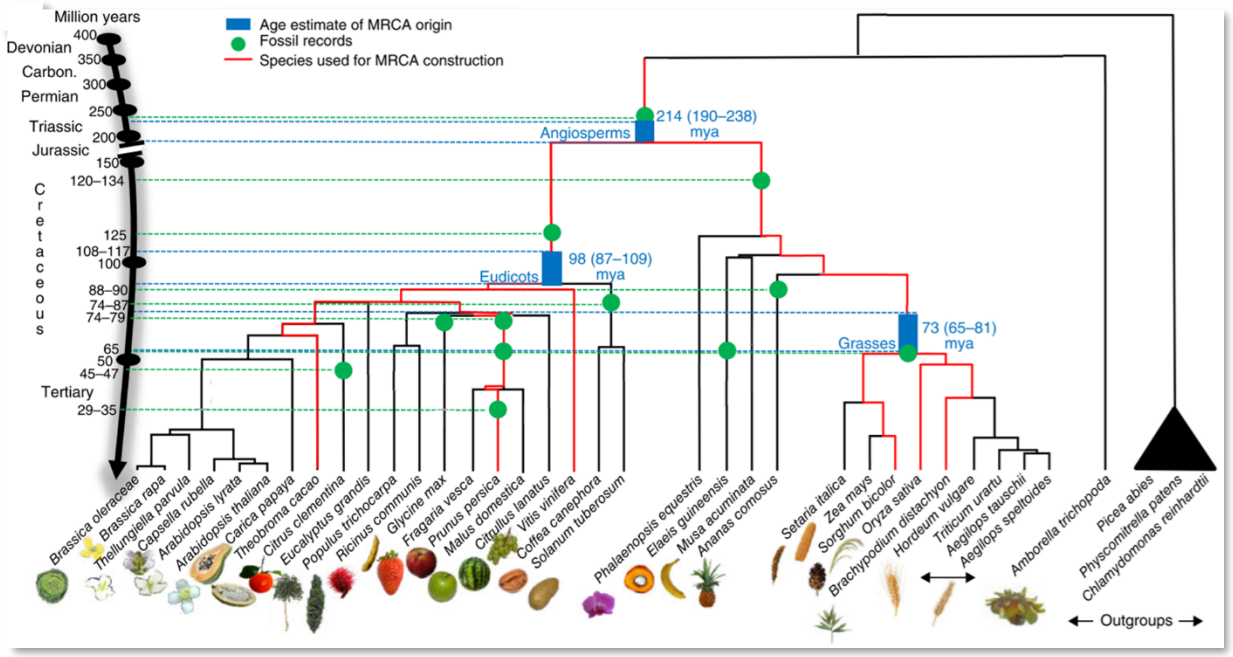


Figure 2 : arbre évolutif des angiospermes d'après F. Murat et al. 2014. Les datations sont mentionnées sur l'échelle des temps à gauche (en millions d'années, MYA) à partir des évidences fossiles répertoriées et matérialisées par un point vert

Cette classification basée sur un critère morphologique arbitraire confère au groupe des dicotylédones un caractère paraphylétique car il ne comprend pas tous les descendants de leur dernier ancêtre commun. En effet, il comprend le groupe des angiospermes basales, dont *Amborella trichopoda*, et celui des eudicotylédones mais exclut celui des monocotylédones alors que la divergence entre ces deux derniers groupes est la plus récente. *A contrario*, la classification basée sur l'étude du nombre d'ouverture dans l'enveloppe du grain de pollen (aperture) regroupe les angiospermes basales et les monocotylédones dont les pollens présentent une unique ouverture alors que ceux des eudicotylédones en comptent trois (plante tricolpate). Cette dichotomie issue de l'observation de deux critères morphologiques distincts illustre la difficulté de l'exercice de la systématique (Wiens, 1995). Au fil des découvertes dans le domaine de la génétique, la prise en compte des caractéristiques des chromosomes et des gènes, désignés par le terme génome créé en 1920 par le botaniste Hans Winkler, a permis de compléter les observations phénotypiques pour rationaliser et préciser la classification des êtres vivants.

1.2 La génomique des plantes à fleur

Le génome est constitué de l'ensemble de l'information génétique présente dans la cellule. La génomique, consiste à étudier sa séquence, son organisation et son fonctionnement. La génomique structurale s'attachera à caractériser l'organisation et la structure des chromosomes, à décrypter la séquence de l'ADN et à annoter, c'est-à-dire identifier et positionner, les gènes et autres séquences caractéristiques tels que les promoteurs ou les éléments transposables. La génomique fonctionnelle a pour objet de comprendre la fonction des gènes et les mécanismes qui régulent leur expression. Les génomes des angiospermes sont particulièrement complexes du fait de leur taille variable et souvent conséquente, de 50 à 125 000 Mb à comparer aux génomes des mammifères de l'ordre de 3000 Mb, de leur fort pourcentage d'éléments répétés, et surtout de la fréquente présence de plusieurs copies de leur matériel génétique initial.

La section suivante du manuscrit a pour objet de montrer comment les progrès technologiques et l'avancée des connaissances accumulées sur différents organismes ont permis de caractériser la séquence d'ADN des génomes des plantes et de comprendre leur structure, puis dans un second temps, de présenter des méthodes disponibles pour explorer leurs fonctionnements.

1.2.1 Le défi du séquençage des génomes complexes

Je retrace ici l'évolution des techniques de séquençage des génomes en me focalisant sur des étapes marquantes, puis présente les différentes méthodes et stratégies qui ont permis de décrypter les génomes complexes des plantes, avec de plus en plus de qualité et d'efficacité.

1.2.1.1 Découverte et premières caractérisations de l'ADN

La molécule d'ADN a été découverte il y a plus de 150 ans, en 1869 par Friedrich Miescher (Dahm, 2008). Ce jeune scientifique suisse cherchait à identifier les composants chimiques présents dans les cellules vivantes. En tentant d'isoler et de catégoriser les protéines et les lipides contenus dans des leucocytes, Miescher a découvert une substance chimique riche en phosphore qu'il a baptisée nucléine du fait de sa localisation dans le noyau cellulaire. En observant la présence de grande quantité de nucléine dans les spermatozoïdes, il a ensuite émis l'hypothèse de son implication dans la fertilité, sans toutefois parvenir à en apporter la preuve. Suite à cette

découverte fondatrice, les avancées se sont succédé. La composition de la nucléine est caractérisée par Albrecht Kossel. Entre 1885 et 1901, Kossel découvre que cette molécule complexe est constituée de deux composants, l'un protéique et l'autre non-protéique, et parvient à isoler et décrire les bases azotées, adénine (A), cytosine (C), guanine (G), thymine (T) et uracile(U), constituant les molécules d'ADN et d'ARN. Ses travaux lui valent le prix Nobel de médecine en 1910 (https://www.nobelprize.org/prizes/medicine/1910/kossel). Plus de 40 années s'écoulent avant qu'en 1944, Avery, MacLeod et McCarthy mettent en évidence que l'acide nucléique est le support de l'information génétique. Pour cela, ils montrent qu'une souche de pneumocoque non-virulente incubée avec l'ADN extrait d'une seconde souche virulente acquiert la virulence (Avery *et al.*, 1944). En 1953, James Watson et Francis Crick décrivent la structure en double hélice de l'ADN (Crick and Watson, 1953) ce qui leur vaudra, associés à Maurice Wilkins, le prix Nobel de physiologie ou médecine en 1962 (https://www.nobelprize.org/prizes/medicine/1962/summary). Cette découverte s'est appuyée sur un cliché de l'ADN par diffraction des rayons X obtenu par Rosalind Franklin. Cependant elle ne fut pas créditée dans la publication de 1953 (Maddox, 2003). En 1958, James Crick émet une hypothèse décrivant la séquence des interactions entre les molécules d'ADN, les molécules d'ARN et les protéines. Il explicitera alors sa théorie, objet de multiples débats, en 1970 dans un article intitulé « *Central Dogma of Molecular Biology* » (Crick, 1970). Dès lors le rôle central de l'ADN dans la biologie étant établi, réussir à le décrypter en le séquençant devient un objectif majeur. Les pionniers Ray Wu et A.D. Kaiser publient la toute première séquence d'un fragment d'ADN de bactériophage (Wu and Kaiser, 1968). Puis, l'Anglais Fred Sanger et l'Américain Walter Gilbert publieront, respectivement avec A. R. Coulson et Allan Maxam, deux méthodes de séquençages de l'ADN (Sanger and Coulson, 1975; Maxam and Gilbert, 1977) et tous deux seront récompensés par le prix Nobel de Chimie en 1980 (https://www.nobelprize.org/prizes/chemistry/1980/summary/). Ainsi débutait, il y a plus de 45 ans, l'histoire des technologies de séquençages (Shendure *et al.*, 2017). La méthode développée par Fred Sanger portera son nom et s'imposera pour constituer la première génération des technologies de séquençages. Les avancées qu'elle aura permises, jusqu'au séquençage du génome humain, ont été relatées par Sanger lui-même dans un article publié en 2001 (Sanger,

2001). Elle est, jusqu'à présent, suivie des méthodes de deuxième et troisième générations qui correspondent à deux ruptures technologiques successives. A l'image de la méthode de Sanger qui s'est imposée par rapport à celle de Gilbert, certaines idées prometteuses n'ont finalement pas porté les fruits attendus alors que d'autres se sont imposées permettant de faire avancer la connaissance du vivant. La présentation qui suit se focalise sur ces dernières.

1.2.1.2 Le séquençage de première génération

Développement de la technologie

La méthode de séquençage Sanger repose sur la synthèse in vitro d'un brin d'ADN complémentaire du brin d'ADN à séquencer, appelé matrice. La réaction de séquençage est initiée par l'hybridation d'une amorce de 10 à 25 nucléotides à la matrice, qui constitue le point de départ de la synthèse du brin d'ADN grâce à l'action d'enzymes, les ADN polymérases, qui recrutent et incorporent les nucléotides complémentaires de ceux du brin matrice. La réaction de synthèse se fait en utilisant deux types de nucléotides, des désoxyribonucléotides (dNTP : désoxyNucléotide TriPhosphate) présents en forte concentration et des didésoxyribonucléotides (ddNTP) disponibles en faible concentration. Les ddNTP se différencient des dNTP par l'absence un groupement -OH, ce qui a pour effet d'interrompre la synthèse du brin d'ADN complémentaire lorsqu'un ddNTP est incorporé. L'incorporation d'un ddNTP étant aléatoire et peu fréquente relativement à celle des dNTP du fait de la différence de représentation, la réaction génère une population de fragments de tailles variables.

C'est sur ce principe que la technologie Sanger s'est développée. Initialement le séquençage d'un fragment d'ADN nécessitait de faire 4 réactions incorporant chacune l'un des 4 nucléotides pour générer une population de fragments pour chacun d'entre eux, puis de faire migrer chaque population sur un gel d'acrylamide pour séparer les fragments en fonction de leur taille au nucléotide près. Les amorces étant marquées, les gels étaient placés sur un film radiographique pour produire l'image à partir de laquelle la séquence était lue, l'alternance des 4 pistes permettant de déduire l'ordre des bases ordonnées en suivant la taille croissante des fragments (Figure 3.A). Cette méthode, dite manuelle, permettait de « lire » des séquences de 200 à 300 nucléotides. La réalisation des réactions de séquençage nécessitait de disposer en quantité

suffisante du brin d'ADN à séquencer. Pour cela, l'ADN de l'organisme à séquencer était fragmenté puis cloné dans un plasmide vecteur et inséré dans un organisme hôte, bactérie ou levure, qui permettait d'amplifier le brin d'ADN initial en se multipliant.

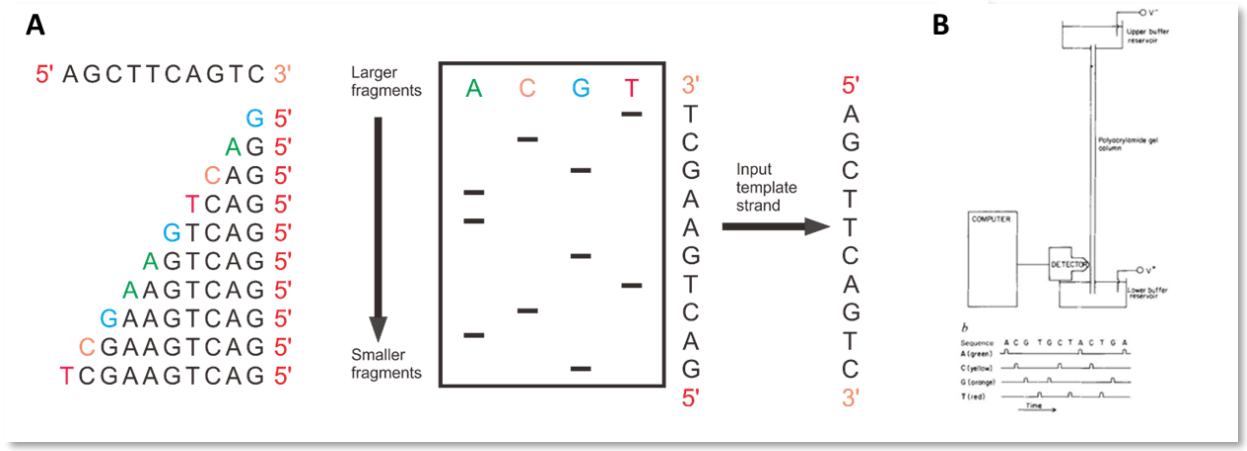


Figure 3 : A. Principe de la méthode de séquençage de Sanger basée sur la reconstruction de la séquence en fonction de la taille de fragments marqués. B. Schéma du premier séquenceur automatique à capillaire publié par Smith et al. en 1986.

Par la suite, la méthode a été progressivement améliorée en simplifiant la réaction de synthèse et l'analyse des fragments. En marquant, non plus les amorces de séquençage mais chacun des 4 ddNTP spécifiquement, une unique réaction de séquençage devient suffisante, au lieu des 4 requises précédemment. La migration des fragments dans des capillaires contenant une colonne de polyacrylamide permet d'éviter les étapes de chargement et de lecture des gels, ce qui ouvre la voie à l'automatisation du séquençage (Smith *et al.*, 1986) (Figure 3.B). Le nucléotide migrant dans la colonne de polyacrylamide est détecté par un tube photomultiplicateur qui capte les signaux émis par les fluorochromes excités par un laser. Ce signal est transcrit sous forme d'une courbe dont les pics représentent les temps de passage des nucléotides devant le système de détection. La séquence des bases nucléotidiques de la molécule d'ADN analysée se présente comme la superposition des quatre courbes de fluorescence et constitue le chromatogramme. Chaque nucléotide est associé à une couleur donnée, à savoir en bleu l'Adénine, en vert la Thymine, en jaune la Guanine et en rouge la Cytosine.

L'apparition des séquenceurs automatiques a constitué une rupture en présentant de nombreux avantages. D'une part, l'automatisation et l'utilisation de la chromatographie au lieu de

l'électrophorèse ont permis un gain de temps considérable. D'autre part, une fois amorti l'investissement lié à l'achat de la machine, le coût de revient à la base chutait drastiquement. Enfin, alors qu'il était difficile de lire plus de 300 nucléotides en employant la méthode manuelle, les séquenceurs automatiques ont permis de produire des lectures de plusieurs centaines de nucléotides, jusqu'à 1000 pour les appareils les plus performants, avec un excellent niveau de qualité, qui demeurera la référence en termes de fiabilité pendant plus de 20 ans. La société Applied Biosystems était partie prenante dans ces innovations et s'est imposée comme le principal constructeur de séquenceur de première génération. Les années 1990 ont vu le développement de centres de séquençages où les séquenceurs automatiques étaient regroupés et les processus étaient industrialisés pour augmenter les débits et réduire les erreurs en standardisant les procédures et en systématisant les contrôles qualité.

Stratégies de séquençage

Parallèlement à l'évolution des technologies de séquençages, différentes stratégies d'assemblage se sont développées. La première stratégie utilisée consistait à séquencer aléatoirement des fragments d'ADN clonés puis à rechercher les chevauchements entre les séquences générées pour constituer des séquences assemblées, ou contigs, selon la méthode de *Shotgun* ou W*hole Genome Shotgun* (WGS) développée par Staden (Staden, 1982). Comme toutes les méthodes de séquençage à ce jour, le WGS implique de disposer des lectures en quantité suffisante pour avoir une surreprésentation du génome à séquencer, qui définit la couverture dont la valeur est calculée en divisant la longueur totale des lectures par la taille estimée du génome. Une couverture élevée permet de multiplier l'occurrence des chevauchements entre lectures et de les confirmer statistiquement, pour en déduire la séquence du génome étudié. Cette méthode applicable aux génomes des virus de 30 à 50 kb a notamment permis d'assembler la séquence du bactériophage lambda dès 1982 (Sanger *et al.*, 1982). Cependant, elle présentait des limites pour séquencer des organismes de tailles supérieures, quelle que soit la couverture utilisée. En effet, proportionnellement à la taille du génome de l'organisme à séquencer, la densité en régions répétées perturbant les logiciels d'assemblage et la probabilité que certaines régions génomiques soient peu ou pas représentées dans les lectures augmente, avec pour conséquence de produire des assemblages fragmentés, sans qu'il y ait de moyens pour ordonner,

ni orienter les contigs les uns par rapport aux autres. En réponse à cette problématique, des stratégies de séquençage visant à établir des ponts entre les contigs se sont développées permettant de décrypter le premier génome bactérien, celui d'*Haemophilus influenzae*, en séquençant les extrémités de fragments d'ADN de 2kb et 16kb sélectionnés aléatoirement dans le génome (Fleischmann *et al.*, 1995). L'information de distance fournit par ces paires de lectures, appelées mate-pairs, étant utilisée pour connecter les contigs générés par l'assemblage des lectures aléatoirement distribuées sur le génome. La possibilité de cloner de grands fragments d'ADN, jusqu'à 150kb, dans un chromosome artificiel de bactérie, BAC pour Bacterial Artificial Chromosome (Shizuya *et al.*, 1992), amplifie le potentiel de cette stratégie qui est théorisée par J.C. Venter (Venter *et al.*, 1996).

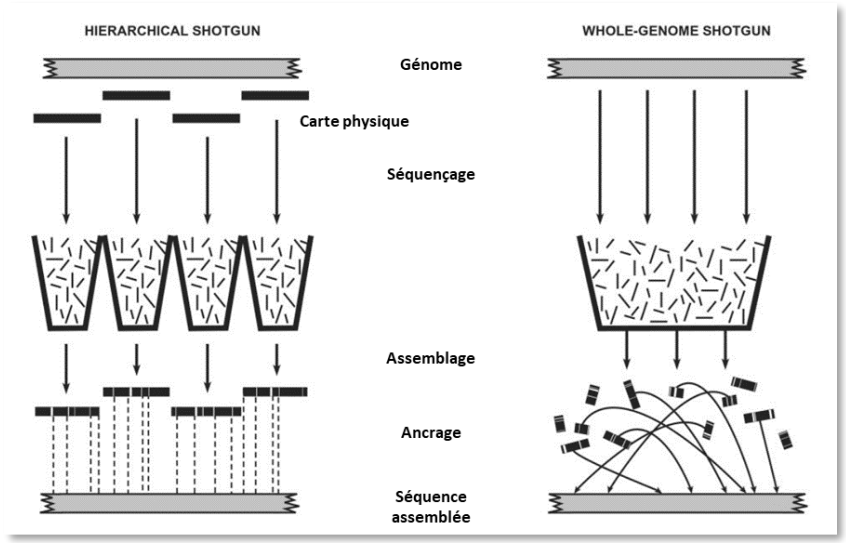


Figure 4 : comparaison des stratégies d'assemblage des lectures. La stratégie de hierarchical shotgun nécessite la construction d'une carte physique qui permettra de réduire la complexité de l'assemblage et d'ancrer les séquences assemblées. La stratégie de whole-genome shotgun se base sur un assemblage sans a priori des lectures. D'après Waterston 2002

A l'opposé de ces stratégies sans a priori sur le choix des fragments d'ADN à séquencer, se développe une stratégie basée sur une organisation des fragments avant leur séquençage, appelée *Hierarchical shotgun* (Figure 4). L'étape préalable au séquençage est de disposer d'une carte physique de grands fragments d'ADN clonés dans des vecteurs BAC ou YAC, *Yeast Artificial Chromosome*, ces derniers ayant l'avantage de pouvoir cloner des fragments d'ADN de plus d'1Mb mais l'inconvénient de présenter des problèmes de stabilité lorsqu'ils étaient utilisés pour

cloner de l'ADN issu d'organismes eucaryotes contenant des séquences répétées. La carte physique est établie par la méthode du Fingerprinting qui consiste à construire une librairie de clones représentant le génome de l'espèce à séquencer, puis à digérer chaque clone par une enzyme de restriction ce qui produit une empreinte constituée par un ensemble de fragments de tailles variées en fonction de la position des sites de restriction reconnus par l'enzyme. La comparaison entre elles de ces empreintes permet de détecter des similitudes traduisant les chevauchements entre clones et, ainsi, de les organiser de proche en proche pour constituer la carte physique du génome (Kohara *et al.*, 1987; Cohen *et al.*, 1993). La carte physique permet de sélectionner un ensemble minimal de clones chevauchants couvrant les chromosomes du génome cible constituant le chemin de recouvrement minimal ou *Minimal Tilling Path* (MTP). Les clones du MTP sont ensuite séquencés et assemblés individuellement selon le principe du séquençage *shotgun* : l'ADN de chaque clone est fragmenté de manière aléatoire en petits morceaux qui sont clonés dans un plasmide et séquencés. Les contigs ainsi générés sont ensuite assemblés sur la base de l'ordre du MTP. Cet assemblage peut ensuite être validé et amélioré en ancrant les contigs sur une carte génétique pour aboutir à un assemblage à l'échelle des chromosomes, les éventuelles séquences manquantes, ou *gaps*, étant comblés par convention par la lettre N dans la séquence de nucléotides.

Organisme	Année de publication	Taille du génome (Mb)	Nombre de gènes	Stratégies de séquençage
Enterobacteria phage λ (virus)	1982	0,048	> 100	WGS
Haemophilus influenzae (bactérie)	1995	1,8	1 700	WGS incluant des lectures Pair-end
Saccharomyces cerevisiae (levure)	1996	12	6 000	Stratégies variées selon les chromosomes
Caenorhabditis elegans * (nématode)	1998	97	18 000	Hierarchical shotgun (Cosmides, YAC) et cartes génétiques
Arabidopsis thaliana (plante)	2000	135	25 000	Hierarchical shotgun (BAC) et cartes génétiques
Drosophila melanogaster (insecte)	2000	165	14 000	Combinaison Hierarchical shotgun (BAC) et WGS
Mus musculus (mammifère)	2002	3400	30 000	Combinaison Hierarchical shotgun (BAC) et WGS

* 1er organisme pluricellulaire

Tableau 1 : Stratégies mise en œuvre pour réaliser les premiers séquençages de génomes.

Par rapport au WGS, le recours à la méthode de *Hierarchical shotgun* implique un travail préparatoire au séquençage plus important, notamment pour la construction de la carte physique, mais permet de simplifier l'assemblage (Waterston *et al.*, 2002) ce qui est crucial pour entreprendre le séquençage des génomes des organismes eucaryotes majoritairement plus grands et plus complexes que ceux des organismes procaryotes. C'est la méthode qui a été choisie pour séquencer le nématode *C. elegans* (The C. elegans Sequencing Consortium, 1998), la plante

modèle *Arabidopsis thaliana* (AGI, 2000) et le riz, *Oryza sativa* (IRGSP, 2005) tandis que des stratégies mixtes reposant principalement sur le WGS ont été utilisées pour séquencer les premiers génomes d'insecte et de mammifère, drosophile (Myers *et al.*, 2000) et souris (Mouse genome sequencing consortium, 2002) (Tableau 1). Au-delà de leurs intérêts propres, l'ensemble de ces travaux allaient servir de prototypes pour réaliser le grand défi des génomiciens de la fin du XX^{e} siècle : le séquençage du génome humain.

Séquençage du génome humain

Dès le début des années 1980, les séquençages de génomes viraux et de la mitochondrie humaine démontrant la possibilité d'assembler une séquence complète à partir des lectures générées par la technologie de Sanger, l'objectif de séquencer le génome humain est devenu une priorité pour la communauté scientifique. Malgré l'étendue de la tâche, le génome humain étant 70 000 fois plus grand que celui du phage λ, des scientifiques du monde entier ont uni leurs forces pour relever ce défi, hors norme à l'époque, motivés par l'enjeu en termes de recherche médicale et la portée symbolique pour l'Homme de séquencer son propre génome. Accompagnant la volonté de cette communauté, la publication des premières cartes génétiques du génome humain (Donis-Keller *et al.*, 1987), la possibilité d'établir des cartes physiques à partir de clones BAC démontrée chez la levure et *C. elegans* (Cohen *et al.*, 1993), et l'automatisation du séquençage, ont constitué des avancées permettant d'élaborer une stratégie basée sur le *Hierarchical Shotgun*. Le projet de séquencer l'ensemble du génome humain est proposé pour la première fois lors de discussions organisées par le ministère américain de l'Énergie au milieu des années 1980, un comité nommé par le conseil national de la recherche américain en approuvant le concept en 1988, tout en soulignant la nécessité de prendre en compte les questions éthiques, juridiques et sociales soulevées par ce projet. C'est sur cette base qu'à la fin de l'année 1990, était créé un projet public soutenu au niveau international, nommé le Projet Génome Humain (*Human Genome Project* HGP), impliquant 20 centres de recherche Nord-américains, Britanniques, Japonais, Français, Allemands et Chinois. Les entreprises de biotechnologie fabriquant les séquenceurs ont également grandement contribué à ce travail en permettant de réduire régulièrement le coût et le temps nécessaire pour générer les séquences d'ADN. Ainsi, le coût du séquençage a été divisé par 100 entre 1991 et 2001, soit une diminution d'un facteur deux tous les 18 mois. Le volume

toujours croissant des données de séquence a entraîné le développement d'outils mathématiques et informatiques pour les assembler et puis les annoter, c'est-à-dire localiser et attribuer des fonctions aux séquences d'ADN (gènes, promoteurs, éléments transposables). La question de la disponibilité et de la pérennité des données a été prise en compte tout au long du projet. Les accords des Bermudes en 1996 et de Fort Lauderdale en 2003 ont été élaborés avec l'idée que la séquence du génome humain devait constituer une ressource communautaire et que les données générées devaient être librement accessibles à tous. Cette notion s'est imposée, bien qu'allant à l'encontre des pratiques alors en vigueur. C'était un enjeu qui allait devenir d'autant plus crucial qu'en 1998 émergeait un projet privé conduit par Craig Venter à la tête de l'entreprise *Celera Genomics* pour poursuivre le même objectif.

Le séquençage du génome, à proprement parler, a débuté en 1995 et se concrétisera 6 ans plus tard par la publication de la version *draft* de la séquence le 15 février 2001 dans Nature (International Human Genome Sequencing Consortium, 2001). Cette séquence couvrait 2,7 Gb soit 81% de la taille totale du génome humain et 94% des régions d'euchromatine. Elle comprenait 87 460 scaffolds assignés aux chromosomes présentant une N50 de 274 kb (taille du scaffold telle que la moitié de la taille totale de l'assemblage soit compris dans des segments de taille supérieure à 274 kb). Son annotation a permis de dénombrer environ 32 000 gènes représentant 5% de la taille totale du génome, à comparer aux 50% occupés par des séquences répétées composées à 90% d'éléments transposables. Cette annotation a donné des précieuses informations sur l'organisation du génome, notamment sur la distribution des gènes, du taux en GC, des îlots CpG ou du taux de recombinaison le long des chromosomes.

L'investissement international entre 1995 et 2001 pour générer les données de séquences est estimé à 300 millions de dollars et 150 millions supplémentaires seront investis dans la production de données complémentaires pour publier la version finale de la séquence portant sa longueur à 2,84 Gb et réduisant le nombre de *gaps* à seulement 341 (International Human Genome Sequencing Consortium, 2004). En agrégeant l'ensemble des sommes investies depuis la genèse du projet HGP, il aura fallu près de 2,7 milliards de dollars pour finaliser la séquence du génome humain (https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost) qui restera le génome le plus complexe décrypté grâce à la méthode Sanger. L'année de la

publication de sa version finale coïncidera avec l'entrée dans l'ère de la deuxième génération des technologies de séquençage qui verra une chute drastique des coûts de production de séquences et l'apparition de nouveaux usages, mais ne sera pas sans effet sur la qualité des génomes produits.

1.2.1.3 Deuxième génération

Dès la fin des années 1980 alors que la technologie Sanger s'imposait en routine dans les centres de séquençage, plusieurs groupes de recherche réfléchissaient à des méthodes alternatives. Leurs travaux porteront finalement leurs fruits 15 ans plus tard, au milieu des années 2000 avec l'apparition des méthodes de séquençage dit de nouvelle génération, *Next Generation Sequencing* (NGS), qui très rapidement supplanteront presque totalement le séquençage Sanger. Le principal changement entre les méthodes de première génération et les technologies NGS réside dans le multiplexage. Ainsi, au lieu de séquencer un fragment d'ADN par tube, le principe des technologies NGS est de séquencer l'ensemble d'une population de fragments d'ADN distincts les uns des autres dans un volume réactionnel unique. De ce fait les technologies NGS sont dites « massivement parallèle », traduction littérale de *massively parallel*. Une population de fragments d'ADN, appelée librairie, est immobilisée sur une surface bidimensionnelle qui permet d'individualiser l'analyse des différents fragments d'ADN tout en utilisant un volume réactionnel unique. Le séquençage proprement dit se fait sur le principe du "séquençage par synthèse" (*Sequencing By Synthesis,* SBS). Il correspond à des cycles successifs de synthèse d'un brin d'ADN par une polymérase, le type de nucléotide incorporé à chaque cycle est déterminé grâce à un système de détection, optique ou électronique selon la méthode mise en œuvre, capable de traiter en parallèle les signaux émis par l'ensemble de fragments d'ADN présents. Pour que le signal émis soit détectable il doit être suffisamment intense, c'est pourquoi chaque fragment d'ADN doit être présent en multiples copies, ce qui implique une étape d'amplification *in vitro* des fragments d'ADN à séquencer. L'une des méthodes d'amplification est le pontage, *bridge amplification*, qui consiste à amplifier les fragments d'ADN grâce à des amorces immobilisées sur une surface solide (Adams and Kron, 1997) afin que les copies générées restent groupées ce qui permet de disposer d'une forte densité de fragment d'ADN sur une surface réduite. L'autre méthode d'amplification consiste à réaliser une amplification par PCR en

émulsion de chaque fragment d'ADN puis d'immobiliser les copies produites sur des billes qui sont ensuite réparties sur une surface percée de plusieurs millions de puits dont les diamètres vont permettre d'isoler une bille par puits (Margulies *et al.*, 2005; Shendure *et al.*, 2005).

A partir de l'amplification, les approches SBS fonctionnent selon deux méthodes pour déterminer la séquence des nucléotides incorporés, soit par additions successives des 4 nucléotides, soit par addition simultanée et terminaison réversible (Goodwin *et al.*, 2016). Le principe des additions successives des 4 nucléotides est de mettre en contact des fragments d'ADN à séquencer un type de dNTP, de détecter s'il est incorporé par la polymérase, de réaliser une étape de lavage, et de recommencer ainsi avec les 3 dNTP restants. Ce cycle complet est ensuite répété tant que la polymérase demeure fonctionnelle. Les approches par addition simultanée et terminaison réversible se définissent par l'utilisation de nucléotides couplés à des molécules terminatrices similaires à celles utilisées dans le séquençage Sanger, dans lesquelles le groupe 3'-OH du ribose est bloqué ce qui empêche l'élongation (Figure 5). Au cours de chaque cycle, un mélange des quatre désoxynucléotides (dNTP) marqués individuellement et bloqués en 3' est ajouté. Après l'incorporation d'un seul type de dNTP, les dNTP non liés sont éliminés et le système de détection identifie quel dNTP a été incorporé. Le groupe bloquant est ensuite retiré et un nouveau cycle peut commencer (Goodwin *et al.*, 2016).

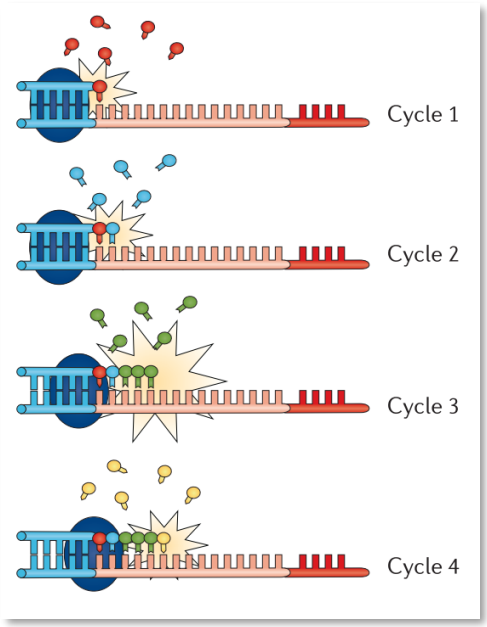


Figure 5 : Principe de séquençage par additions de nucléotide : un seul type de nucléotide est présent à chaque cycle. L'incorporation de plusieurs dNTP identiques lors d'un cycle se traduit par une augmentation du signal émis. D'après Goodwin et al. 2016.

En complément de la méthode d'amplification et d'immobilisation des fragments d'ADN et d'incorporation des nucléotides, les méthodes de détection de l'incorporation des dNTP caractérisent les trois principales méthodes de séquençage SBS. Le pyroséquençage de Ronaghi et Nyrèn repose sur la méthode des additions successives des 4 nucléotides. Si le dNTP est incorporé, un pyrophosphate est libéré et convertie en adénosine-triphosphate (ATP) par une sulfurylase en présence de PAPS (3'-Phosphoadenosine-5'-phosphosulfate), puis l'ATP est lui-même converti en bioluminescence par une réaction catalysée par la luciférase en présence de luciférine. La lumière émise est détectée par le séquenceur (Ronaghi *et al.*, 1996). Une deuxième approche également basée sur les additions successives des quatre nucléotides consiste à enregistrer l'incorporation de dNTP grâce à la détection des ions naturellement émis par leur polymérisation avec un transistor à effet de champs (Purushothaman *et al.*, 2006). Enfin, la troisième approche utilisant l'addition simultanée et la terminaison réversible, et qui s'est largement imposée dans le temps, consiste à générer un signal correspondant à l'incorporation par la polymérase des dNTP grâce à des dNTP marqués par fluorescence (Mitra *et al.*, 2003).

Les premières plateformes NGS intégrées sont apparues en 2005, la société 454 proposant un séquenceur appliquant les méthodes de PCR en émulsion, additions successives et pyroséquençage (Margulies *et al.*, 2005) , tandis que le séquenceur de la société Solexa s'appuie sur les méthodes d'amplification par pontage, addition simultanée et terminaison réversible, et détection de la fluorescence. Ces séquenceurs constituent une révolution en termes d'accès au séquençage pour les laboratoires. Alors qu'au terme du projet séquençage du génome humain, le séquençage à grande échelle était encore l'apanage de quelques centres de génomique, grâce au séquenceur 454 et aux autres instruments concurrents qui ont suivi de près, les laboratoires individuels ont pu accéder instantanément à une capacité équivalente à celle d'un centre de génomique de l'ère du HGP. Cette "démocratisation" de la capacité de séquençage a eu un impact profond sur la culture et la composition du domaine de la génomique, de nouvelles méthodes, de nouveaux résultats, de nouveaux génomes et d'autres innovations apparaissant de toutes parts. Les débits ont continué à progresser, ainsi en 2020 un séquenceur Illumina NovaSeq pouvait générer en deux jours et pour quelques milliers d'euros plus d'un milliard de lectures indépendantes, totalisant un terabase de séquence, soit 40 fois la totalité des séquences

générées (23 gigabases) pendant l'ensemble du projet international de séquençage du génome humain.

Contrairement au monopole d'Applied Biosystems sur le séquençage Sanger, plusieurs sociétés, dont 454 (rachetée par Roche), Solexa (rachetée par Illumina), Agencourt (rachetée par Applied Biosystems), Helicos, Complete Genomics et Ion Torrent, se sont livrées à une concurrence intense qui a entraîné une évolution rapide du paysage avec de nouveaux instruments présentés de façon régulière proposant des progrès constants. Entre 2007 et 2012, le coût brut, par base, du séquençage de l'ADN a chuté de quatre ordres de grandeur (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data). A partir de 2012, le rythme d'amélioration ralenti et la concurrence se réduisent, la société Illumina affirmant sa position dominante bien que la solution Complete Genomics acquise par la société chinoise MGI appartenant au *Beijing Genomics Institute* (BGI) et de nouveaux acteurs tels que Elements Biosciences et Ultima Genomics se positionnent dès 2022 comme concurrents potentiels.

Application au séquençage de génome de novo

Bien que révolutionnaires notamment en termes de volume de données générées, ces séquenceurs présentaient leurs propres limitations. En 2012, les séquences produites avaient encore des taux d'erreur plus élevés, jusqu'à 15% des bases identifiées, et des longueurs de lecture généralement plus courtes (de 35 à 700 pb selon les technologies) que celles obtenues avec la méthode Sanger (Liu *et al.*, 2012). Cependant, des méthodes d'analyse se sont développées très rapidement pour prendre en compte ces spécificités et notamment la courte longueur des lectures (Whiteford *et al.*, 2005), qui pouvait être compensée par l'utilisation de séquences appariées (Bashir *et al.*, 2008). Ces séquences appariées, désignées par les termes *mate-pairs* ou *paired-ends* selon les protocoles utilisés pour les produire, correspondent à deux courtes séquences dont la distance les séparant (de 2 à 20kb) et leurs orientations relatives sont connues. Elles permettent lors de l'assemblage d'une part de mieux prendre en compte les régions répétées et d'autre part de relier et d'orienter les contigs les uns par rapport aux autres créant ainsi des scaffolds (Figure 6).

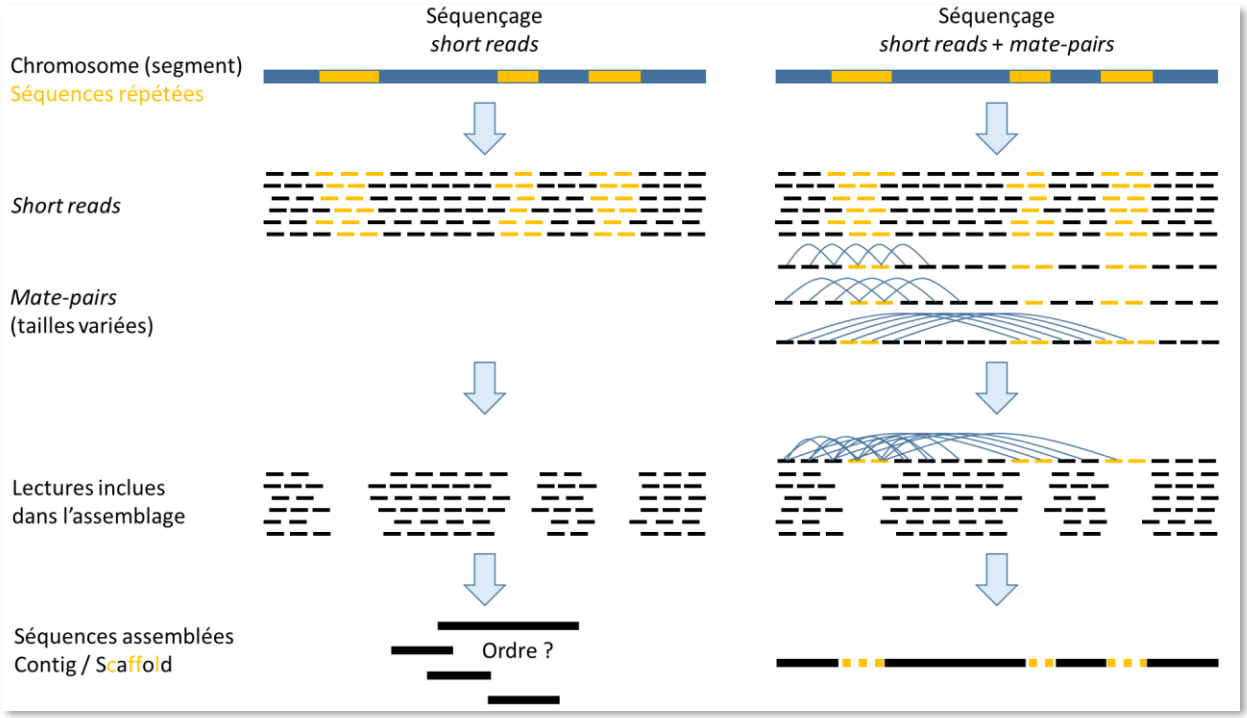


Figure 6 : Intérêt de l'utilisation des séquences appariées, mate-pairs, pour l'assemblage en comparaison d'un assemblage basé uniquement sur des short reads isolés.

A ce jour, les lectures, bien qu'elles soient toujours plus courtes que celles du séquençage Sanger sont exactes à plus 99,9 %.

Avec l'avènement du NGS en 2005, le nombre d'assemblages *de novo* a considérablement augmenté (Michael and Jackson, 2013; Kersey, 2019). L'incompatibilité *a priori* entre des lectures courtes et des génomes très répétés a été surmontée par des algorithmes d'assemblage dédiés, basés sur les graphes de Bruijn, tel que EULER (Pevzner *et al.*, 2001) puis Velvet (Zerbino and Birney, 2008). Néanmoins, les assemblages produits par ces logiciels étaient assez médiocres lorsqu'ils étaient utilisés pour assembler de grands génomes (>1 Gb), en comparaison avec les résultats obtenus dans le cadre du projet génome humain.

Deux métriques principales sont utilisées pour évaluer la qualité d'un assemblage. La N50 qui correspond à la taille du segment (contig ou scaffold) telle que la moitié de la taille totale de l'assemblage soit compris dans des segments de taille égale ou supérieure. La L50 qui correspond au plus petit nombre de segments (contigs ou scaffolds) dont la somme des tailles représente la moitié de la taille estimée du génome. Ces deux métriques sont anticorrélés, la L50 déclinant

lorsque la N50 augmente et inversement. Le couple de valeurs, forte N50 et faible L50, indique que l'assemblage est continu et donc de bonne qualité.

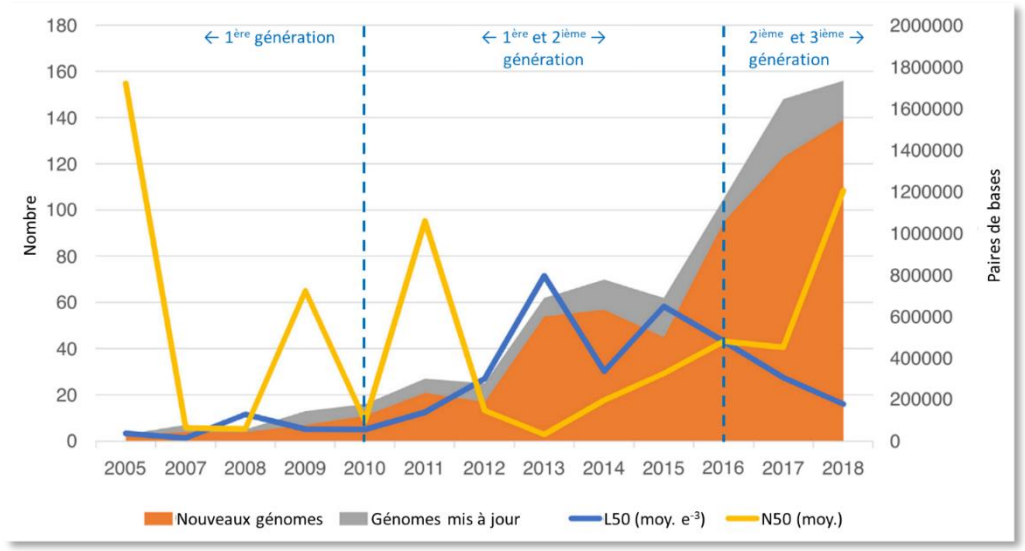


Figure 7 : Évolution du nombre et de la qualité des assemblages de génomes de plantes soumis à l'INSDC (réseau des bases de données DDBJ, EMBL-EBI et NCBI), par an, entre 2005 et 2018 (les statistiques pour 2018 couvrent la période de janvier à août). Les zones orange et grises indiquent le nombre d'assemblages, respectivement nouveaux et mis à jour (axe de gauche). Les lignes bleues et jaunes indiquent les L50 et N50 moyennes en paires de bases (axe de droite). Figure adaptée de Micheal & Jackson 2013 et de Kersey 2019.

Bien que les longueurs de lecture plus courtes soient en partie à l'origine de la baisse de qualité (Figure 7), leur impact est généralement surestimé (Wang *et al.*, 2021). L'autre facteur auquel imputer ce recul en termes de qualité est le choix fait dans le cadre des nombreux projets de séquençage de ne pas investir dans les méthodes permettant d'organiser les contigs entre eux pour produire de grands scaffolds, également appelé pseudo-molécules. Alors que les génomes humains, de la souris ou d'*Arabidopsis* s'appuyaient sur des cartes physiques et des cartes génétiques, ces dernières ont été délaissées au profit de stratégie se basant uniquement sur le séquençage shotgun et les lectures appariées. Cependant des méthodes innovantes, cartes optiques et analyse de la conformation des chromosomes, allaient émerger à partir de 2014 et offrir des solutions nouvelles pour retrouver des génomes de hautes qualités à partir de 2016, après 5 années de recul dans la qualité moyenne des génomes déposés dans les bases de données (Figure 7).

Application au reséquençage de génomes

En complément du séquençage de génome *de novo*, les NGS ont ouvert la voie au reséquençage de génomes à grande échelle avec l'objectif d'identifier les bases génétiques des variations phénotypiques *via* des études d'association entre variations génétiques et caractères d'intérêt (*genome-wide association studies*, GWAS) (Shendure *et al.*, 2017). Le reséquençage consiste à aligner les *shorts reads* (*mapping*) sur un génome de référence et à identifier les différences existantes, par exemple entre deux génotypes de plantes ou au sein de deux populations humaines. C'est une tâche très différente de l'assemblage de génome *de novo*. Là où pour assembler les séquences en contig, l'assemblage *de novo* implique de rechercher le maximum d'homologie de séquences entre des lectures chevauchantes, aux erreurs attendues près, le *mapping* des séquences pour le reséquençage doit tolérer des absences ponctuelles d'homologie qui correspondent à la réalité biologique (le polymorphisme entre individus d'une espèce). L'élément discriminant pour faire cette distinction est la profondeur de séquençage. La profondeur de séquençage correspond au nombre de lectures obtenues indépendamment pour chaque base ciblée et s'exprime en nombre de fois (x). Une profondeur de 100x signifie que la base, ou la région considérée, est couverte par 100 lectures indépendantes (Lacoste *et al.*, 2017). La profondeur minimum nécessaire pour identifier un polymorphisme nucléotidique, *single nucleotide polymorphism* (SNP), varie selon la plante considérée, en fonction de la complexité de la structure de son génome et notamment de son niveau de ploïdie (cf. section suivante du manuscrit), allant d'un minimum de 8x pour une plante diploïde à un minimum de 15x pour une plante tétraploïde (Clevenger *et al.*, 2015).

Les projets de reséquençage se sont largement développés dans le domaine de la génétique humaine avec un nombre croissant de génomes reséquencés dans le cadre d'initiatives extrêmement ambitieuses telles que le projet 100 000 génomes (Barwell *et al.*, 2018). Chez les plantes les travaux de reséquençage ont débuté avec la plante modèle *Arabidopsis thaliana* (Cao *et al.*, 2011) et sur le riz (Huang *et al.*, 2013). Ensuite, ils se sont développés pour des plantes aux génomes plus complexes *via* des approches qui portaient, soit, sur les génomes complets comme pour l'étude des résistances à la sécheresse chez le maïs (Xu *et al.*, 2014), soit, sur des régions ciblées, par exemple pour étudier les processus de domestication de l'orge (Pankin *et al.*, 2018).

1.2.1.4 Troisième génération

Au début des années 2010, alors que les technologies NGS étaient des technologies largement utilisées, dont les avantages (débit, coûts, qualité des lectures) et les inconvénients (courtes lectures, amplifications des matrices induisant des erreurs) étaient bien connus, deux méthodes de séquençage en temps réel de molécules uniques permettant d'obtenir de longues lectures, *long reads*, ont émergé.

Séquençage SMRT (PacBio)

La première approche a été lancée par Webb et Craighead (Levene *et al.*, 2003) puis développée par Korlach et Turner au sein de la société Pacific Biosciences (PacBio) (Eid *et al.*, 2009). La technologie, baptisée *single molecule real time sequencing* (SMRT), est basée sur la fixation d'une polymérase au fond d'un micro-puits où se produit la synthèse d'un brin d'ADN complémentaire à partir d'une molécule d'ADN simple brin circulaire (Figure 8.A). Lors de la réaction de séquençage, la polymérase incorpore des dNTP marqués par fluorescence, ce qui génère un signal lumineux détecté par une caméra. Les micro-puits, appelés *zero-mode waveguide* (ZMW), constituent des dispositifs nano-photoniques car leur diamètre est inférieur à la longueur d'onde de la lumière émise. Cette propriété permet de canaliser le signal lumineux produit et d'éviter les interférences entre les 8 millions de micro-puits adjacents présents sur le support de séquençage, nommée SMRTCell. La synthèse continue jusqu'au moment où la polymérase s'arrête spontanément (après 10 000 à 100 000 nucléotides incorporés) générant de très longues lectures. Le taux d'erreur est relativement important (environ 10 %) mais celles-ci survenant de façon aléatoire, il est possible de les corriger en « lisant » plusieurs fois le brin d'ADN circulaire. Les lectures obtenues par cette méthode, *circular consensus sequencing* (CCS), ont des valeurs de fiabilité comparables à celles de la technologie Illumina (Logsdon *et al.*, 2020). Par ailleurs, le SMRT permet de détecter directement la méthylation des nucléotides, s'appuyant sur le fait que le temps d'incorporation d'un nucléotide méthylé diffère de celui d'un nucléotide non modifié.

Séquençage nanopore (ONT)

Une deuxième approche est le séquençage Nanopore. Ce concept repose sur le passage d'une molécule d'ADN simple brin dans un pore de quelques nanomètres de diamètre sous l'effet d'un champ électrique, provoquant l'émission de flux d'ions spécifiques du type de nucléotide (Figure 8.B). Ainsi, il est possible d'établir la séquence des nucléotides traversant le pore en mesurant ces flux ioniques (Deamer *et al.*, 2016). Le principe était connu et considéré comme une piste pour le futur du séquençage de l'ADN depuis la fin des années 1980, mais des décennies de travail seront nécessaires pour concrétiser le concept et construire un séquenceur efficace. Le principal problème résidait dans le fait que le passage du brin d'ADN à travers le nanopore était extrêmement rapide et le nombre d'ions par nucléotide très faible, les deux phénomènes rendant trop imprécise la mesure des courants ioniques pour en déduire une séquence. Pour résoudre ce problème, il a notamment fallu mieux caractériser, puis modifier les protéines du nanopore et développer des méthodes pour améliorer l'analyse des signaux résultants (Branton *et al.*, 2008). Ces avancées se sont concrétisées par la création de la société *Oxford Nanopore Technologies* (ONT) par Bayley en 2005. En 2014, le MinIon, premier séquenceur basé sur les propriétés du Nanopore est commercialisé par ONT. Il produit des lectures de longueurs au moins égales et parfois largement supérieures aux lectures générées par la méthode de PacBio. En effet, la taille de la molécule à séquencer est la seule limite théorique au séquençage nanopore et des lectures atteignant 2000 kb ont été obtenues (Payne *et al.*, 2018). Le principal handicap de la technologie est la distribution non aléatoire des erreurs qui empêche une correction par la profondeur de lecture ou le séquençage redondant des brins d'ADN. Les lectures nanopore doivent encore être corrigées par des *shorts reads* de haute qualité (Illumina par exemple). Cependant les propriétés des pores et les méthodes d'analyses du signal progressent constamment et les taux d'erreurs inhérents à la technologie baissent régulièrement, passant de 40% en 2014 à 10% en 2018 (Rang *et al.*, 2018), et continuant à se réduire régulièrement depuis. Une différence majeure par rapport aux autres technologies de séquençage est la possibilité de miniaturiser les dispositifs de nanopores, qui peuvent être aussi petits qu'une clé USB, car ils reposent sur la détection de signaux électroniques et non optiques.

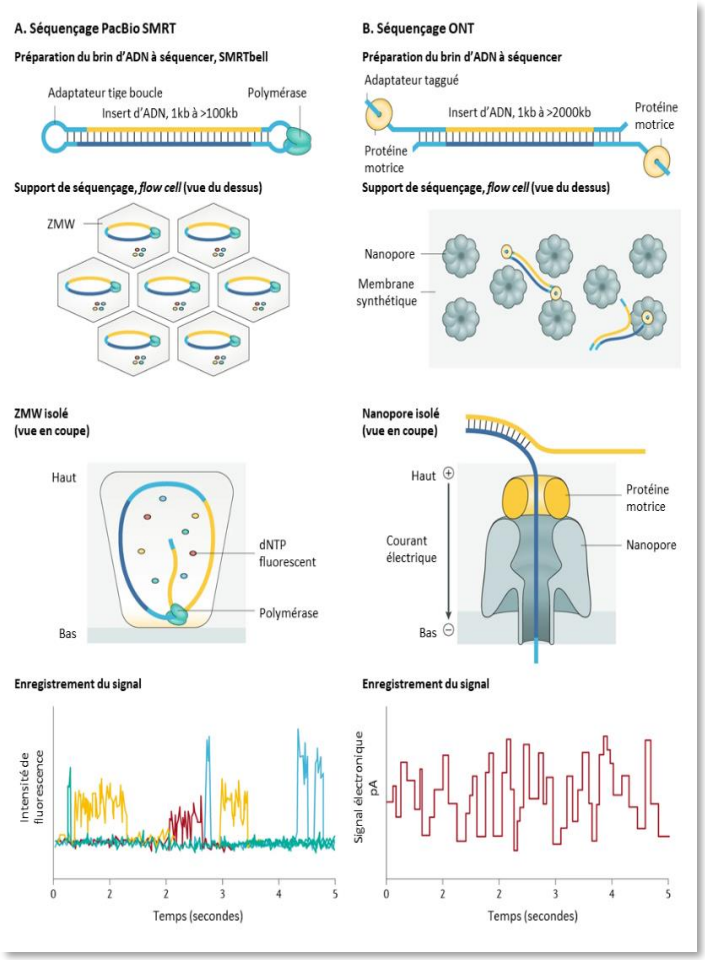


Figure 8 : Présentation des technologies de séquençage de troisième génération, dites long reads. A. Séquençage SMRT (PacBio). L'ADN (jaune : brin sens, bleu : brin antisens) est fragmenté et lié à des adaptateurs tige boucle (bleu clair) pour former une molécule circulaire appelée SMRTbell. La SMRTbell est couplée à une ADN polymérase et chargée sur une SMRTcell pour le séquençage. Chaque SMRTcell contient jusqu'à 8 millions de micro-puits de quelques picolitres (ZMW). La lumière pénètre dans les 20-30 nm inférieurs de chaque puits, correspondant à un volume de détection de 20 zl (10^{-21} l). Lors du chargement une SMRTbell et la polymérase associée s'immobilisent au fond de chaque ZMW. Des dNTP fluorescents sont ajoutés pour initier le séquençage. Lorsque la polymérase commence à synthétiser le nouveau brin d'ADN, un dNTP fluorescent est brièvement maintenu dans le volume de détection et une impulsion lumineuse appliquée sur le fond du puits excite le fluorophore. La lumière ainsi émise par le fluorophore est détectée par une caméra, qui enregistre la longueur d'onde et la position relative du signal sur la SMRTcell. Le fluorophore lié au phosphate est ensuite clivé du nucléotide lors de l'incorporation de la base dans le nouveau brin d'ADN et libéré dans le tampon, empêchant toute interférence fluorescente pendant l'impulsion lumineuse suivante. La séquence d'ADN est déterminée par la succession des émissions des fluorophores qui est enregistrée dans chaque ZMW, une couleur différente correspondant à chaque base d'ADN (par exemple, vert, T ; jaune, C ; rouge, G ; bleu, A). B. Séquençage Oxford Nanopore Technologies (ONT). L'ADN natif (jaune : brin sens, bleu : brin antisens) est lié à des adaptateurs de séquençage (bleu clair) couplés à une protéine motrice sur une ou deux extrémités. Ce complexe est chargé sur la flowcell pour le séquençage. Celle-ci porte des milliers de nanopores auxquels les molécules d'ADN se lient. L'adaptateur de séquençage s'insère dans l'ouverture du nanopore puis la protéine motrice déroule l'ADN double brin. Un courant électrique est appliqué et fait passer l'ADN chargé négativement à travers le nanopore à une vitesse d'environ 450 bases par seconde provoquant des perturbations caractéristiques du courant, ce qui génère un signal, squiggle, correspondant à un k-mer particulier (c'est-à-dire une chaîne de bases d'ADN de longueur k). L'analyse de la suite des k-mer permet de déduire la séquence d'ADN.

Le séquençage des premiers génomes de plantes, Arabidopsis et le riz, en 2000 et 2005 respectivement, basé sur l'approche BAC et le séquençage Sanger, a produit des assemblages de grande qualité qui figurent encore aujourd'hui parmi les meilleurs assemblages de génomes végétaux. L'introduction de la technologie de séquençage Illumina a permis de séquencer les génomes en grands nombres mais, souvent, avec une faible contiguïté. Ces assemblages présentent généralement un catalogue de gènes relativement complet et correctement assemblé, mais les régions répétées et notamment riches en éléments transposables sont fragmentées, voire sous-représentées. Ces lacunes compliquent drastiquement l'étude de la dynamique des éléments transposables et celle des interactions à distance entre gènes et séquences régulatrices. La maturité croissante des technologies de séquençage long reads, concrétisée par la mise sur le marché des séquenceurs ONT et PacBio, a changé radicalement la donne permettant des assemblages de haute qualité. En 2018, J.M. Aury et son équipe du Génoscope ont étudié la qualité des assemblages de 105 génomes végétaux au regard des technologies utilisées (Belser *et al.*, 2018). Ce travail avait deux conclusions notables. La première conclusion était qu'à cette époque seulement 10% des génomes de plantes séquencés possédaient un assemblage avec un N50 > 5 Mb, dont le riz, Arabidopsis et *Brachypodium* séquencés en Sanger. La seconde conclusion était la prédominance des séquençages par les deux technologies long reads parmi les génomes de haute qualité. Depuis 2018, la tendance s'est confirmée avec l'avènement de nouvelles méthodes de scaffolding (présentées ci-après), une meilleure intégration des données et la baisse des taux d'erreurs des technologies *long reads*, concrétisées par le CCS de PacBio, permettant de générer de longues lectures de hautes qualités. En 2021, 15 espèces végétales de génomes de tailles variées ont été séquencées avec la technologie PacBio CCS dans le cadre de projets conduits au CNRGV. Tous les assemblages obtenus, dès lors que la profondeur de séquençage atteignait 20x, égalaient ou surpassaient les assemblages répertoriés dans l'étude publiée en 2018 par le Génoscope (Figure 9), confirmant la conclusion de Logsdon qui indiquait en 2020 que « si le séquençage est le 'microscope' avec lequel les généticiens étudient la variation génétique, et il est clair que les technologies *Pacbio HiFi* et *ONT ultralong reads* fournissent une nouvelle lentille et un nouvel objectif pour comprendre la variation, la structure et l'organisation de l'ADN et de l'ARN » (Logsdon *et al.*, 2020).

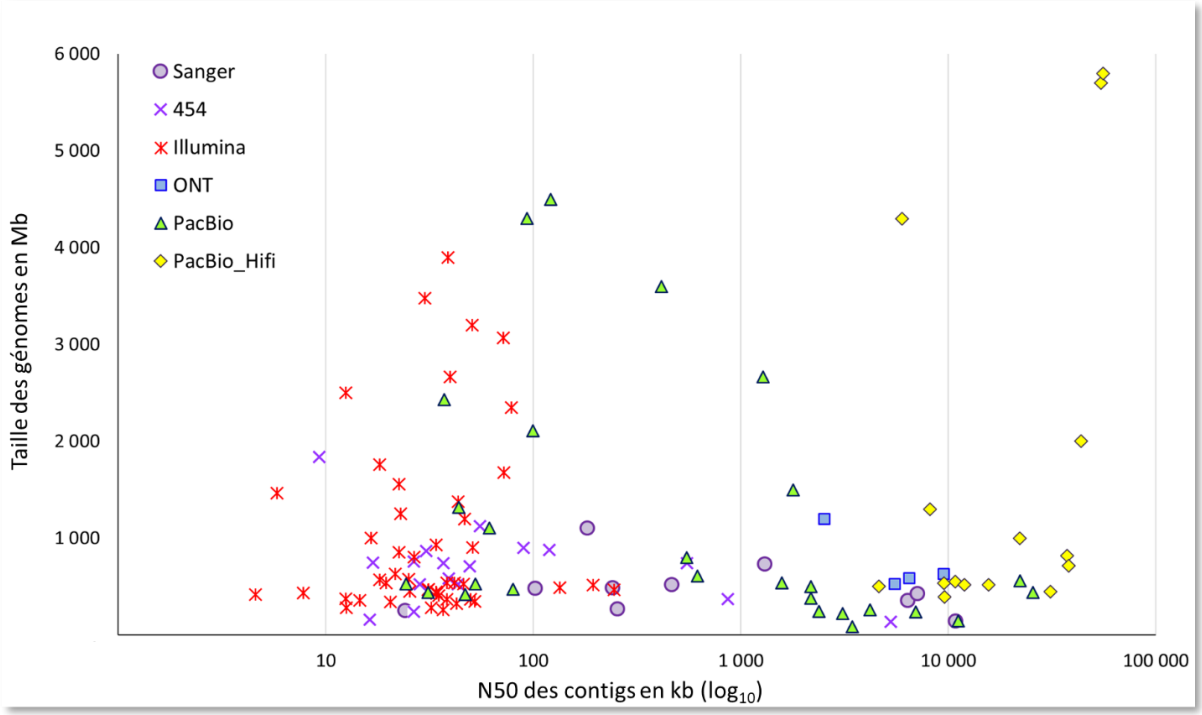


Figure 9 : Comparaison du N50 des contigs et de la taille des génomes de 120 assemblages de génomes de plantes en fonction de la principale technologie de séquençage utilisée : Sanger (10), 454 (17), Illumina (47), ONT (4), PACBIO (27) et PacBio Hifi (15). Adapté de Belser et al. 2018 et complété par les données internes au CNRGV de génomes séquencés par la technologie PacBio CCS.

1.2.1.5 Technologies de « scaffolding »

Malgré les progrès permis par les technologies long reads, l'assemblage seul des lectures, dans la plupart des cas, ne permet pas d'obtenir la séquence continue de tous les chromosomes d'un organisme eucaryote. C'est particulièrement vrai dans les cas des plantes qui se caractérisent par des génomes de grandes tailles avec de forts taux de répétitions en comparaison, notamment, des vertébrés (Jiao and Schneeberger, 2017) (Figure 10).

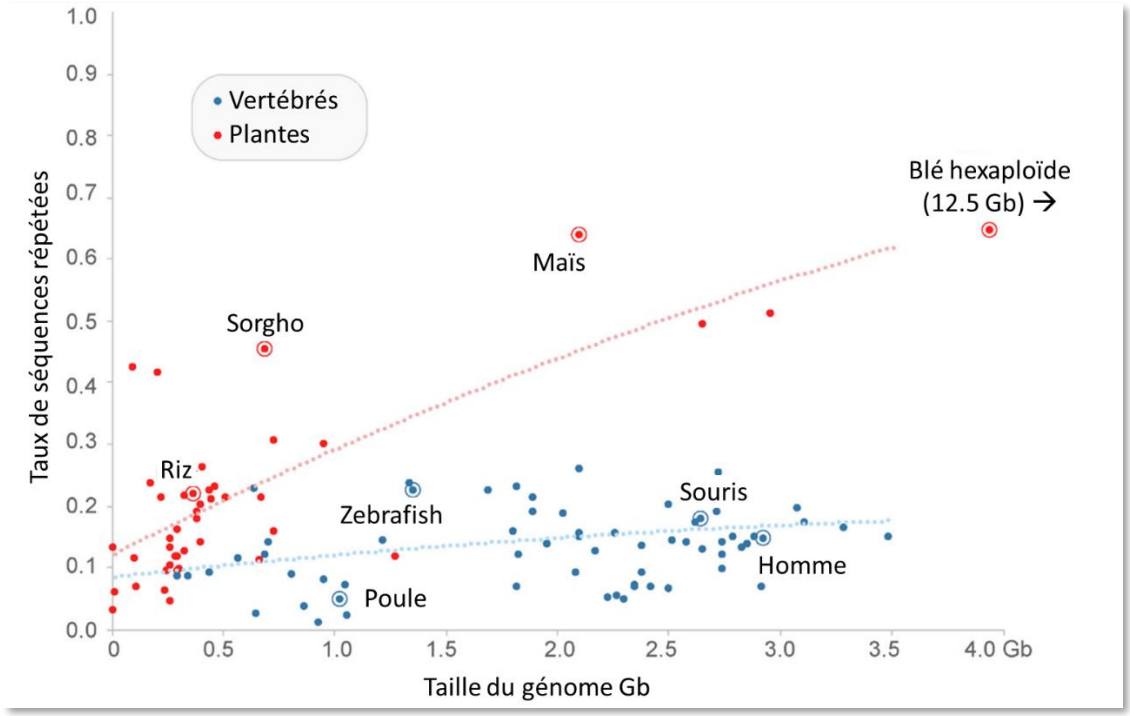


Figure 10 : Comparaison de la taille et du taux de séquences répétées entre des génomes de plantes et de vertébrés. 44 assemblages de génomes de plantes et 68 assemblages de génomes de vertébrés sont examinés en termes de taux de séquences répétées et de taille du génome. Les génomes végétaux présentent des tailles et des taux de séquences répétées globalement supérieurs à ceux des vertébrés et la présence des séquences répétées chez les plantes est davantage corrélée à la taille du génome que chez les vertébrés. Adapté de Jiao et Schneeberger, 2017

Pour tendre vers des assemblages complets et continus, dit telomere-to-telomere ou gold-standard, il est nécessaire d'ordonner et d'orienter les contigs pour produire des scaffolds les plus grands possibles correspondant idéalement à des chromosomes entiers. Les premières solutions pour faire ce travail étaient les cartes physiques d'inserts clonés et les cartes génétiques. Elles sont aujourd'hui, respectivement, remplacées et complétées par les méthodes d'analyses de conformation des chromosomes et par les cartes optiques.

Méthodes d'analyse de conformation des chromosomes

Ces méthodes d'assemblage à l'échelle du chromosome sont basées sur la méthode Hi-C, développée à l'origine pour étudier le repliement tridimensionnel du génome et les interactions entre les chromosomes (Lieberman-aiden *et al.*, 2009). La méthode Hi-C est basée sur la création de liens physiques artificiels entre fragments d'ADN qui sont physiquement proches dans la conformation naturelle des chromosomes au sein du noyau. Ils sont liés *in situ* avant d'être extraits puis clivés par des enzymes de restriction. Les fragments d'ADN proches se retrouvent

liés, formant une paire de séquences, et sont séquencés ensemble en utilisant une technologie *short reads*. Statistiquement la majorité des paires de lectures générées à partir des deux extrémités de ces fragments proviennent de loci proches l'un de l'autre et ces paires de lecture peuvent donc être utilisées pour scaffolder l'assemblage (Selvaraj *et al.*, 2013). Ainsi, cette stratégie permet d'établir des liens entre des loci distants de 10 à 100 Kb. La méthode est améliorée grâce à une modification du protocole Hi-C, baptisé Chicago (Dovetail Genomics). Partant du principe que les interactions de longue distance entre chromosomes dans les noyaux, tels que les associations fréquentes entre les télomères de différents chromosomes, brouillent les signaux utiles pour l'assemblage, la méthode Chicago vise à les supprimer pour ne conserver que les signaux imputables aux liaisons physiques intra-chromosomiques. Pour cela, le protocole mis en œuvre consiste à reconstituer la chromatine *in vitro* à partir d'ADN purifié puis à capturer les contacts chromatiniens (Putnam *et al.*, 2016). La combinaison de ces données avec les assemblages des séquences produit des scaffolds présentant des N50 sensiblement améliorés par rapport aux valeurs des séquences seules (https://dovetailgenomics.com/dovetail-tree-of-life/) et constitue une stratégie couramment mise en œuvre désormais pour le séquençage des génomes végétaux (Jarvis *et al.*, 2017; Jiao *et al.*, 2017; Lightfoot *et al.*, 2017).

Cartes optiques

La technologie des cartes optiques consiste à observer des fragments d'ADN marqués par fluorescence à des sites définis pour établir des empreintes, ou cartes, des brins d'ADN. Cette méthode peut être considérée comme un développement des *fingerprints* de clones pour produire les cartes physiques utilisées dans les projets de séquençage par *hierarchical shotgun*. De la même façon que les fingerprints des clones sont alignés sur la base des recouvrements pour construire des MTP, les cartes individuelles sont assemblées pour produire des cartes consensus à l'échelle du génome, qui seront utilisées pour scaffolder les contigs issus d'un assemblage de séquences correspondant. La méthode de cartographie optique a été inventée au début des années 1990 (Schwartz *et al.*, 1993), mais il n'est devenu possible de la mettre efficacement en œuvre qu'avec une somme d'améliorations continues des systèmes (Yuan *et al.*, 2020) aboutissant à la commercialisation de systèmes à haut débit Irys, puis Saphyr, par la société BioNano Genomics. Ces systèmes se basent sur un faisceau de nano-canaux qui permettent de

linéariser efficacement les molécules d'ADN marquées pour les faire migrer devant une caméra qui les détecte et les analyse en continu (Ernest *et al.*, 2012). La taille requise pour les fragments analysés est de 150 kb minimum, ce qui impose de maîtriser les protocoles d'extraction d'ADN de haut poids moléculaire. L'assemblage des cartes issues de ces grands fragments permet de produire des cartes consensus de tailles supérieures à plusieurs Mb. La combinaison des données de séquençage et des cartes optiques fonctionne particulièrement bien. En effet, les cartes optiques, grâce à la taille des molécules analysées qui excède celle de la plupart des régions répétées, offrent une image précise de celles-ci alors que ces régions complexes causent typiquement des cassures dans les assemblages basés uniquement sur les lectures NGS. Cette complémentarité est illustrée par la publication de génomes de plantes de très haute qualité assemblés *telomere-to-telomere* pour la majorité des chromosomes tel que le génome du lupin blanc (Hufnagel *et al.*, 2020).

L'intégration des données de cartes optiques comme celle des données de conformation des chromosomes suivent une approche en deux étapes. Tout d'abord les cartes optiques consensus ou les données de conformation sont alignées sur les contigs. Les éventuels conflits sont détectés et résolus en cassant les contigs. Les contigs obtenus sont ensuite utilisés dans un second temps pour produire les scaffolds sur la base des alignements. L'intégration combinée des données de conformation et de la cartographie optique est susceptible d'améliorer encore la contiguïté et surtout la validité des assemblages en vérifiant la convergence des résultats obtenus par ces deux méthodes basées sur des concepts différents (Jiao *et al.*, 2017).

1.2.1.6 L'intérêt de combiner les technologies de séquençage : l'exemple du blé.

Parallèlement au développement des technologies de séquençage de 2^{ième} et 3^{ième} générations, s'est couru un long marathon pour séquencer le génome du blé hexaploïde. Le blé hexaploïde est, à la fois, une plante d'intérêt majeur fournissant 20% des calories consommées par la population mondiale et, un Everest de complexité, composé de trois sous-génomes homéologues (A, B et D) représentant une taille totale de 15 Gb, près de 40 fois la taille du génome du riz. Ainsi, le séquençage du génome du blé était une nécessité autant qu'un challenge pour les génomiciens du début des années 2000. Lors de la formation du consortium international de séquençage du génome du blé (IWGSC) en 2005, le premier constat fut la double impossibilité,

économique et technique, d'appliquer les méthodes de carte physique à l'ensemble du génome et d'utiliser la technologie Sanger pour produire les séquences. La stratégie adoptée fut de partager les efforts en utilisant les ressources cytogénétiques disponibles pour le cultivar de blé *Chinese spring* (CS) permettant d'individualiser les chromosomes par cytométrie (Šafář *et al.*, 2004) pour distribuer aux 20 pays membres du consortium des chromosomes ou des bras de chromosomes. Chaque pays avait la charge d'appliquer la même stratégie basée sur le séquençage de clones BAC organisés en MTP. Cependant, même en réduisant la complexité à un chromosome, le travail demeurait colossal et la première séquence, celle du chromosome 3B, a été publiée 9 ans après le lancement du projet (Choulet *et al.*, 2014). Parallèlement, l'ampleur de la tâche conduisait certains chercheurs à choisir une autre voie consistant à séquencer les génomes d'espèces diploïdes apparentées telles que *Aegilops tauschii*, travaillant ainsi sur des génomes 3 fois plus petit, sans devoir prendre en compte les relations d'homéologies existantes entre les 3 sous-génomes au sein du génome du blé hexaploïde. Mais là encore, même en mettant en œuvre des méthodes à haut débit pour produire et ancrer les MTP de BAC, le travail demeurait fastidieux et les résultats émergeaient lentement (Luo *et al.*, 2013). L'éclaircie apparaît au début des années 2010 avec l'avènement du séquençage short reads à haut débit. Le séquenceur Illumina HiSeq 2000 capable de produire à chaque run 200 Gb de lectures appariées de 100 paires de bases, offrait la possibilité d'assembler les régions faiblement répétées des génomes de blé diploïde en mettant en œuvre une stratégie de séquençage shotgun. Ainsi furent publiés des assemblages des blés diploïdes *Aegilops tauschii* et *Triticum urartu,* progéniteurs des génomes A et D (Ling *et al.*, 2013; Jia *et al.*, 2013). Mais ces assemblages présentant des N50 des contigs très faibles, respectivement, 3,42 et 4,51 kb, ne pouvaient être considérés que comme des catalogues de gènes, néanmoins très utiles pour étudier des gènes d'intérêts et développer des marqueurs moléculaires.

Si les lectures inférieures à 100 pb des technologies Illumina seules semblaient rédhibitoires pour travailler sur le blé hexaploïde, leur utilisation combinée avec celles produites par la technologie 454 atteignant 500 pb et les données des génomes des blés diploïdes ont permis d'obtenir un premier assemblage du génome du blé *Chinese spring* permettant d'identifier entre 94 000 et 96 000 gènes dont les deux tiers étaient assignés à l'un des trois sous-génomes A, B et D (Brenchley

et al., 2012). Du point de vue méthodologique cette publication mettait en évidence l'intérêt de développer des stratégies intégrant plusieurs types de données de séquences complémentaires. Dans le même temps l'IWGSC complétait sa stratégie initiale par le séquençage par la technologie Illumina (séquences appariées) d'ADN isolé à partir de chromosomes purifiés. Cette stratégie permit d'obtenir un assemblage de 10,2 Gb soit 61% du génome du blé présentant des N50 des contigs de 1,7 à 8,9 kb selon les chromosomes. 133 090 gènes *high confidence* ont été annotés sur cette séquence et 56% d'entre eux ont été assignés à l'un des bras des 21 chromosomes. Cette séquence identifiée en tant que *chromosome survey sequence* (CSS) fut mise à disposition de la communauté scientifique en 2014. Immédiatement après cet important jalon, l'année 2015 constitue un tournant avec le développement des technologies de séquençage *long reads* de PacBio, l'avènement des technologies de *scaffolding*, Hi-C et cartographie optique, et, surtout, le développement de nouveaux algorithmes pour assembler les génomes complexes de manière beaucoup plus rapide et précise. En 2017, grâce à ces innovations couplées à l'utilisation de MaSuRCA, un nouveau pipeline développé pour prendre en compte à la fois les lectures courtes et longues, paraissaient coup sur coup des assemblages des génomes de blé diploïde puis hexaploïde présentant des N50 de contigs de 486,6 et 232,6 kb, nettement améliorées au regard des assemblages des blés diploïdes parus précédemment (Zimin, Puiu, Luo, *et al.*, 2017; Zimin, Puiu, Hall, *et al.*, 2017). Parallèlement paraît un assemblage annoté du génome de CS, baptisé TGACv1, basé sur des lectures courtes, *paired-ends* et *mate-pairs*, analysés par des logiciels dédiés. Les N50 des contigs et des scaffolds sont relativement faibles, 16,7 et 83,9 kb, mais cet assemblage préfigure le développement de nouveaux logiciels d'analyse dédiés.

C'est ainsi que les années 2017-2018 marquent un tournant décisif dans le séquençage du génome du blé. Un premier génome de référence d'un blé polyploïde, le génome du blé tétraploïde *Triticum dicoccoides* est séquencé, assemblé et publié en l'espace de quelques années seulement à l'aide du logiciel développé par la société NRGene (Avni *et al.*, 2017). Le génome est assemblé à partir de 5 librairies de séquences short reads Illumina appariées présentant des tailles d'inserts de 450 pb à 10 kb et représentant une profondeur de lecture totale de 176x. Le logiciel DeNovoMAGIC2 développé par NRGene est utilisé pour assembler les contigs et faire une première étape de scaffolding, complétée dans un second temps par l'intégration des données

d'une carte génétique dense et de données Hi-C. L'assemblage final de 10,5 Gb, soit 87,5 % de la taille estimée du génome présentait une N50 des contigs de 57 kb et de près de 7 Mb pour les scaffolds. Sur la base de cette preuve de concept impressionnante et avec tous les outils disponibles, l'assemblage de haute qualité de génome du blé hexaploïde devenait possible. Ce fut le cas en en août 2018, l'IWGSC publiant le génome de référence RefSeq v1.0 après 13 ans de travail. L'assemblage final s'est appuyé sur un assemblage de courtes lectures appariées par le logiciel DeNovoMAGIC2 complété par toutes les données utiles générées au fil du projet : cartes physiques, données de génotypage par séquençage, cartes optiques Bionano et données Hi-C (Appels *et al.*, 2018). L'assemblage final compte 21 pseudomolécules correspondant aux 21 chromosomes, sa taille totale est de 14,5 Gb. Les N50 des contigs, scaffolds et super scaffolds sont, respectivement, de 52 kb, 7 Mb et 22,8 Mb.

Depuis 2018, de nouveaux génomes de blés hexaploïdes ont été séquencés en s'appuyant, soit, sur les courtes lectures pour produire 10 génomes (Walkowiak *et al.*, 2020), soit, sur les longues lectures ONT (Aury *et al.*, 2022). La technologie Hifi de PacBio, qui a fait ses preuves sur l'orge (Mascher *et al.*, 2021) constitue également une alternative crédible pour séquencer des génomes aussi complexes que celui du blé.

L'histoire du séquençage du génome du blé est un parfait exemple de la façon dont les choses avancent dans le paysage de la génomique. Les ruptures technologiques (ici le séquençage *short reads* haut débit, les outils de scaffolding, de nouvelles méthodes informatiques, les longues lectures) imposent de pouvoir réajuster la stratégie, tout en gardant à l'esprit l'utilité de méthodes plus anciennes telles que les cartes génétiques ou les méthodes de tri de cellules et des chromosomes pour disposer d'ADN de haute qualité et ainsi maximiser les résultats des protocoles de séquençage longues lectures, et des méthodes de *scaffolding*, cartes optiques ou Hi-C. Le génome du blé a également valeur d'exemple pour démontrer qu'en combinant les méthodes adéquates il est possible d'obtenir des assemblages complets, continus et exacts à l'échelle des nucléotides ce qui constitue le « *gold standard* » recherché pour mener l'ensemble des applications possibles à partir d'un génome de référence (établir les liens entre génotype et phénotype pour des caractères d'intérêt, développer de marqueurs, étudier l'évolution à toutes les échelles du SNP au changement du caryotype).

La dynamique constante de l'évolution des technologies rend l'exercice de définir une stratégie « idéale » pour produire un génome *gold standard* pour une espèce donnée, à la fois complexe et empreint d'une part de subjectivité. Cependant à ce jour, la stratégie qui tend à s'imposer consiste à combiner des lectures avec un minimum d'erreurs, de longues lectures et des méthodes de scaffolding indépendantes pour disposer d'informations complémentaires et permettant une validation croisée des résultats (Jung *et al.*, 2020). Dans cette optique, l'utilisation des lectures PacBio Hifi qui répond aux deux premiers impératifs et la combinaison d'une carte optique Bionano couplée à une l'analyse de conformation Omni-C fait figure, à ce jour, de solution optimale en termes de qualités des données tout en en limitant les ressources informatiques nécessaires pour l'assemblage et la complexité de celui-ci.

La prochaine étape sera d'atteindre le « *platinum standard* » intégrant en plus des caractéristiques précitées l'haplotypage de l'ensemble du génome. De tels génomes commencent à être publiés et se révèlent utiles pour étudier l'évolution des espèces (Zhang *et al.*, 2021), la transmission des caractères (Q., Zhou *et al.*, 2020) ou les bases de l'hétérosis (Jan *et al.*, 2019). Dans le futur les coûts de séquençage continueront à baisser permettant, pour chaque espèce, de disposer de plusieurs génotypes séquencés en fonction des caractères prioritairement étudiés ou pour représenter la diversité. Une étude portant sur 100 variétés de tomates séquencées avec la technologie ONT a permis de démontrer les liens entre les variations structurales et l'expression des gènes, notamment pour des caractères d'intérêt agronomique (Alonge *et al.*, 2020). Pour tirer parti de ces données il faudra, parallèlement aux capacités de séquençage et d'assemblage, développer les outils d'annotation, notamment pour détecter les relations d'orthologies entre gènes, et de détection des variations structurales (Golicz *et al.*, 2016).

A la suite de cet historique des technologies de séquençages et d'assemblage des génomes (dont la publication en article de revue sera considérée), la suite de cette section du manuscrit est consacrée aux connaissances majeures acquises sur l'organisation des génomes de plantes à partir de l'exploitation de ces ressources

1.2.2 Structure des génomes végétaux angiospermes

L'essor du séquençage des génomes s'est traduit par la publication d'un nombre croissant de génomes couvrant de mieux en mieux la diversité inter- et intraspécifique des plantes terrestres (Figure 11).

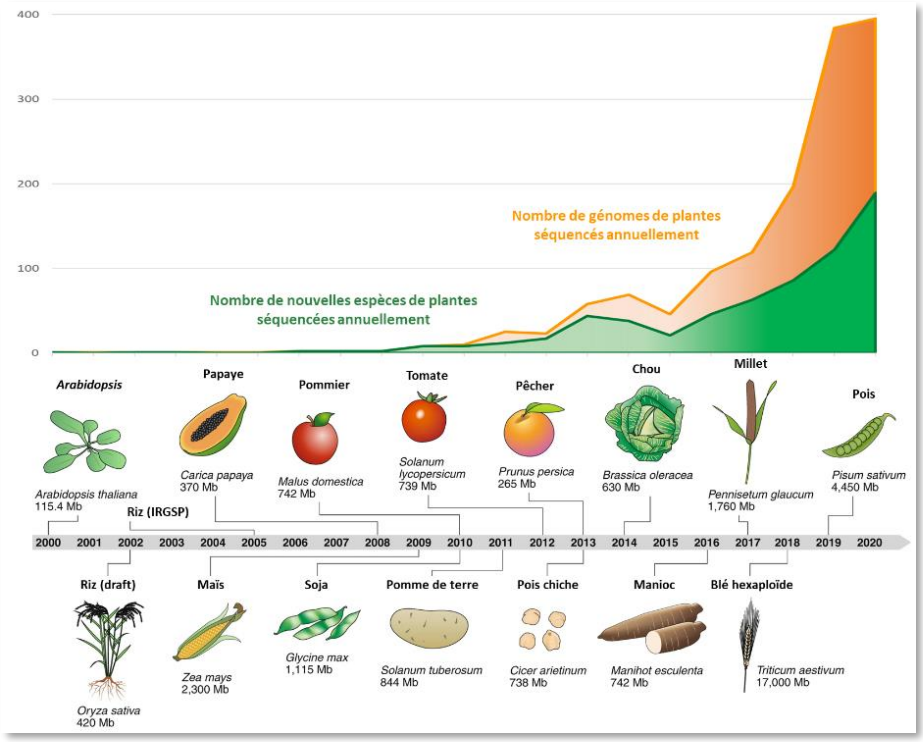


Figure 11 : Historique du Séquençage des plantes à fleurs. En haut : nombre de publications de génomes de plantes, en orange, et de nouvelles espèces séquencées, en vert (données : https://www.ncbi.nlm.nih.gov/genome). En bas : année de la publication de la séquence du génome et la taille estimée du génome. A. thaliana est présenté car il constitue le premier génome végétal séquencé, les autres espèces ont été choisies pour leur importance agronomique. D'après Purugganan et Jackson, 2021.

La comparaison des caractéristiques des génomes publiés en masse a permis de mettre en évidence une grande diversité de structure des génomes au sein des angiospermes (Liu *et al.*, 2015). La structure d'un génome est définie par un ensemble de valeurs telles que la taille, le nombre de chromosomes, le nombre de gènes et les caractéristiques de ceux-ci, et la densité en éléments transposables (ET). A titre d'exemple, le génome d'*Arabidopsis thaliana* est comporte 5 chromosomes pour 120 Mb et 20% d'ET, alors que celui de *Zea Mays* (maïs) comporte 10 chromosomes pour 2 Gb et 74% d'ET (Tableau 2).

Espèce	Sm	Am	Pe	Mu	Os	Bd	Si	Sb	Zm	At	Gr	Pb	Vv
Nombre chromosome (n)	27	13	19	11	12	5	9	10	10	5	13	17	19
Taille génome (Mb)	212.80	706.30	1086.20	473.00	382.80	271.90	400.10	738.50	2058.60	119.70	775.20	509.00	486.20
nombre (transcrit) gène	22,285	26,846	29,431	36,536	33,127	26,413	38,125	27,159	78,483	26,637	40,976	42,341	25,329
Longueur gène (Mb)	36.63	151.50	283.85	131.34	72.40	75.61	115.82	79.90	183.15	50.87	101.86	117.94	155.24
Séquence codante (Mb)	25.53	25.40	26.42	37.97	33.45	33.98	41.58	34.25	62.16	33.10	45.25	56.99	29.82
Nombre intron	100,572	82,937	86,301	161,132	95,376	106,747	124,169	104,612	188,877	112,745	144,200	180,965	129,570
Longueur intron totale (Mb)	11.14	126.69	252.19	93.37	38.95	41.63	54.79	45.66	121.00	17.76	48.90	60.95	125.42
Longueur intron moyenne (bp)	111	1,528	2,922	579	408	390	441	436	641	158	339	337	968
Longueur TE dans les introns (bp)	2.81	46.74	125.58	2.55	1.37	2.85	11.04	8.80	40.02	0.46	0.87	6.16	47.22
TE total (Mb)	102.15	425.66	667.90	31.90	97.99	77.43	186.03	435.33	1524.76	23.87	79.43	190.44	239.96
Proportion TE totale (%)	48.00	60.30	61.50	6.70	25.60	28.50	46.50	58.90	74.10	19.90	10.20	37.40	49.40
Proportion TE dans introns (%)	25.17	36.89	49.80	2.73	3.53	6.84	20.15	19.27	33.07	2.56	1.78	10.11	37.65
			Monocotylédones							Dicotylédones			

Tableau 2 : Illustration de la diversité structurale des angiospermes par les caractéristiques les plus communément utilisées pour décrire et comparer les génomes. Am, Amborella tichopoda ; At, Arabidopsis thaliana ; Bd, Brachypodium distachyon ; Gr, Gossypium raimondii ; Si, Setaria italica ; Mu, Musa acuminata ; Os, Oryza sativa ; Pe, Phalaenopsis equestris ; Pb, Pyrus brestchnediri ; Sb, Sorghum bicolor ; Sm, Selaginella moellendorffii ; Vv, Vitis vininfera ; Zm, Zea mays. TE, éléments transposables. Adapté de Cai et al. 2014.

Les nombres de gènes sont, par contre, relativement homogènes, variant entre 25 329 et 42 341 au sein du panel, à l'exception notable du maïs qui présente le chiffre de 48 483 gènes annotés, conséquence directe d'un évènement de duplication complète de son génome survenu il y a 5 millions d'années.

1.2.3 Au-delà de la structure : analyse de la régulation des génomes de plantes

Pour comprendre comment le génome d'une plante contrôle son phénotype, la connaissance de sa séquence doit être complétée par des données décrivant l'expression spatiale et temporelle des gènes, et les facteurs susceptibles de la réguler, tel que l'état épigénétique. Dans ces domaines, les technologies et les méthodes d'analyses ont fortement évolué et tendent à d'avantage d'exhaustivité et de qualité. L'objectif ici n'est pas de faire un inventaire exhaustif de technologies en perpétuelle évolution mais de dresser un bref historique des méthodes d'analyse et de présenter celles qui seront utilisées dans l'étude présentée dans la deuxième partie de ce manuscrit.

1.2.3.1 Analyse de l'expression

Comme pour l'assemblage des séquences des génomes, les méthodes de séquençage de nouvelle génération ont constitué une révolution pour l'étude de l'expression des gènes en permettant de séquencer les transcriptomes, c'est-à-dire l'ensemble des ARN messagers issus de l'expression d'un génome. Les premières tentatives d'analyse de la transcription des gènes remontent aux années 1980 avec le séquençage par la méthode Sanger des *expressed sequence*

tags (EST), qui sont de courtes séquences nucléotidiques générées à partir de la rétrotranscription des ARN en ADNc (Adams *et al.*, 1991). Puis, au milieu des années 1990, deux approches différentes à l'échelle génomique ont émergé pour caractériser les transcrits, à savoir l'analyse sérielle de l'expression des gènes (SAGE) (Velculescu *et al.*, 1995) et les puces à ADN (Lockhart D. J. *et al.*, 1996). La méthode SAGE implique le séquençage de longs concaténaires de petites séquences appelées étiquettes (initialement ∼10 pb) qui identifient de façon unique différents ARNm. Une analyse statistique de la fréquence des étiquettes permet une quantification directe des transcriptions et la découverte de nouveaux gènes. La méthode a progressivement évolué pour identifier plus précisément les étiquettes en portant leur longueur à 17, 21 puis 26 paires de bases (Matsumura *et al.*, 2005). L'autre méthode, les puces à ADN, ou *microarrays*, est basée sur l'hybridation des brins marqués d'ADNc de l'échantillon étudié avec les sondes immobilisées sur une surface solide bidimensionnelle (Schena *et al.*, 1995). En raison de leur haut débit et de leur faible coût, les microarrays ont été largement utilisés dans les années 2000. Cependant, contrairement à SAGE, les puces à ADN impliquent de disposer d'un génome ou d'un transcriptome de référence ce qui limite les capacités de la méthode qui ne permettra pas de détecter des transcrits correspondants à des gènes non annotés dans le génome de référence ou absents du transcriptome de référence.

Le séquençage massif des transcrits par les technologies NGS, dit RNAseq, va remédier aux limites inhérentes aux approches précédentes. En effet le RNAseq permet à la fois de découvrir de nouveaux transcrits et de quantifier l'expression de l'ensemble du transcriptome. Un génome ou un transcriptome de référence est fréquemment utilisé pour aligner les lectures, mais si ces données de référence ne sont pas disponibles, un transcriptome peut être assemblé *de novo* à l'aide des séquences générées et être utilisé pour quantifier les transcrits séquencés en les alignant sur cette nouvelle référence. Au-delà de la quantification de l'expression des gènes, la technique de RNAseq est très efficace pour détecter les variations de la séquence des transcrits issus de l'épissage alternatif des ARNm pré-messagers. Ainsi, le RNAseq est devenu la méthode de choix pour l'analyse du transcriptome depuis une décennie (Wang *et al.*, 2019). Cependant, reposant sur des lectures courtes issues des technologies NGS à haut débit (Illumina principalement), le RNAseq présente plusieurs limitations dont la principale est l'impossibilité de

distinguer les uns des autres les multiples variants d'épissage en tentant d'en reconstituer la séquence pleine longueur à partir de courtes lectures (Steijger *et al.*, 2013). Ce problème était particulièrement prégnant dans le cas des génomes complexes des organismes eucaryotes qui présentent un grand nombre d'isoformes par gène et dont les gènes ont plusieurs promoteurs candidats et extrémités 3' (Conesa *et al.*, 2016). La technologie de séquençage de longues lectures de PacBio permet de résoudre ce problème par le séquençage pleine longueur et haute qualité des isoformes issues de l'épissage, grâce à la méthode de séquençage IsoSeq (Chao *et al.*, 2018).

1.2.3.2 Analyse d'une modification épigénétique majeure : la méthylation

L'épigénétique désigne l'étude des facteurs qui modifient l'expression des gènes sans modifier la séquence nucléotidique de l'ADN. Les marques épigénétiques sont transmissibles entre générations mais, cependant, réversibles. Elles affectent la molécule d'ADN soit au niveau des histones, ce qui a un impact sur le niveau de condensation de la chromatine, soit directement au niveau de l'ADN. Les principales modifications des histones sont l'acétylation, la méthylation, la phosphorylation et l'ubiquitination. Les différentes combinaisons de ces modifications impactent l'état chromatinien permettant ou, au contraire, restreignant l'activité des protéines régulatrices de la transcription (Jenuwein and Allis, 2001). Les marques épigénétiques majeures de l'ADN correspondent à la méthylation des cytosines. Chez les plantes, elle a lieu dans 3 contextes spécifiques notés CG, CHG et CHH (avec H = A, C ou T). La méthylation est catalysée par des ADN méthyltransférases et guidée par de petits ARN interférents de 24 nucléotides dans le cadre d'un processus cellulaire désigné par le terme de RdDM, *RNA-directed DNA methylation* (Law and Jacobsen, 2011). La méthylation de l'ADN affecte différentes régions de l'ADN aboutissant à des impacts divers. La méthylation des séquences répétées, notamment des éléments transposables (ET), tend à supprimer leur transcription et leur prolifération prévenant ainsi les effets délétères qui peuvent être associés à leur insertion (Lisch, 2009). La méthylation des gènes se produit principalement dans le contexte CG, au niveau des promoteurs et des régions transcrites, *gene-body methylation*, et est associée à la transcription des gènes (Zilberman *et al.*, 2007). Les technologies NGS permettent d'étudier l'état de méthylation des cytosines à l'échelle du génome entier avec une résolution au niveau de la paire de bases grâce à la méthode de séquençage bisulfite. Cette méthode consiste à traiter l'ADN génomique par le bisulfite de

sodium qui a la propriété de convertir les cytosines non méthylées en uraciles, transformant la marque épigénétique en variation génétique identifiable par séquençage (Yong *et al.*, 2016). Cette méthode est actuellement la méthode de choix pour les analyses de méthylation mais elle pourrait être supplantée dans l'avenir par les technologies de séquençage de troisième génération, capables de déterminer l'état de méthylation natif des nucléotides (Flusberg *et al.*, 2010).

1.2.3.3 Intégration de donnés multiomiques

Les données présentées précédemment traduisent les niveaux de mutation (par séquençage de l'ADN), d'expression (par séquençage de l'ARN) et de méthylation (par séquençage bisulfite) associés aux gènes et à leurs promoteurs. Elles font partie de la famille des données omiques. C'est ainsi que sont désignées les données biologiques acquises grâce aux technologies d'analyse omiques : génomique, transcriptomique, protéomique et métabolomique. Ces technologies ont pour objet de mesurer quantitativement les facteurs qui gouvernent l'expression des gènes : variations de leur séquence, état épigénétique, régulation de la transcription et conséquences sur la traduction. L'analyse combinée de données omiques hétérogènes a pour objectif de détecter des corrélations entre tout ou partie des données, tout ou partie des espèces, génotypes, individus, et tout ou partie des gènes considérés. Ces corrélations peuvent traduire des relations fonctionnelles entre les différents niveaux de régulation examinés. L'analyse deux à deux des données omiques (mutation / expression, expression / méthylation, …) se fait efficacement, mais constitue une approche orientée qui ne donne qu'une vision partielle de la multiplicité des liens fonctionnels à l'œuvre dans la cellule. L'enjeu est donc d'analyser simultanément des ensembles de données omiques hétérogènes pour cumuler un ensemble d'informations complémentaires, afin d'obtenir une vue d'ensemble plus complète du système biologique. C'est l'objet de l'intégration omique, ou intégration multiomique (Cantini *et al.*, 2021).

Plusieurs familles de méthodes permettent de réaliser de l'intégration omique : méthodes de réduction de dimension, méthodes utilisant des modèles probabilistes dites bayésiennes, méthodes basées sur la similarité

La réduction de dimension, ou réduction de la dimensionnalité, consiste à sélectionner ou à créer un nombre restreint de variables, appelées variables latentes, qui conservent la majorité des informations contenues par l'ensemble initial des variables, par nature de grande dimension. La création des variables latentes est généralement très efficace mais présente l'inconvénient majeur d'être difficile à interpréter car la nouvelle variable peut ne pas être directement reliée à une donnée biologique ce qui complique la compréhension des processus biologiques sous-jacents. Cependant, cette famille de méthodes est très utilisée pour intégrer des données omiques et comprend des logiciels largement utilisés, en particulier mixOmics (Rohart *et al.*, 2017).

Les méthodes bayésiennes consistent à calculer les probabilités de diverses causes hypothétiques à partir de l'observation d'événements connus. Elles nécessitent que les données présentent des distributions probabilistes. Pour le vérifier, une hypothèse a priori sur les distributions des données est faite, puis en observant celles-ci les distributions sont ajustées. Le choix des distributions se base sur la connaissance des processus biologiques de l'utilisateur des méthodes. Cela présente l'avantage de faciliter l'interprétation biologique des résultats mais s'avère complexe et, surtout, de nature à influencer assez significativement les résultats a posteriori (Chauvel *et al.*, 2020).

Les méthodes basées sur la similarité reposent sur l'estimation de la similarité ou de la distance entre les variables. En fonction des valeurs de similarité les données sont classées et pour chaque type de données omiques un réseau de similarité est construit. Ces réseaux sont ensuite fusionnés pour produire un réseau de similarité multiomique utilisé pour faire l'intégration omique des données. A l'instar des méthodes de réduction de dimension, les méthodes basées sur la similarité produisent des résultats potentiellement difficiles à interpréter d'un point de vue biologique.

Les méthodes d'intégration multiomique sont largement utilisées en médecine humaine, par exemple pour l'étude du cancer, mais peu mises en œuvre pour l'étude des forces à l'œuvre au cours de l'évolution des génomes des plantes ; en octobre 2022 les mots-clés « multi-omics polyploidy », « multi-omics plant genome » et « multi-omics cancer » recherchés sur le site de référence pubmed.gov identifient, respectivement, 7, 197 et 2225 publications.

2 La génomique comparée et la paléogénomique

Les méthodes de séquençage et d'analyse de la régulation des génomes ont permis de bâtir un socle de connaissances complémentaires pour étudier la biologie des plantes. Au cours de ce travail, ces connaissances ont été utilisées pour disséquer les mécanismes qui régissent l'évolution des génomes. La comparaison des caractéristiques génomiques des espèces a permis d'identifier les évènements évolutifs qui ont façonné leurs génomes. Pour appréhender les méthodes de comparaison mises en œuvre, il est nécessaire de présenter deux domaines de recherche : la génomique comparée et la paléogénomique.

2.1 La génomique comparée : établir les liens évolutifs entre les génomes

La génomique comparée met en relation les génomes d'espèces distinctes en identifiant des homologies de séquences entre des régions chromosomiques sur la base de repères communs. La conservation de l'ordre de ces repères au sein de blocs chromosomiques définit la colinéarité, ou synténie. Il est à noter que cette définition ne correspond pas au sens initial du terme. Celui-ci avait été introduit pour la première fois pour désigner des *loci* localisés sur le même chromosome chez l'homme (Renwick, 1971), mais sa signification a dévié de cette définition initiale pour acquérir son sens actuel (Passarge *et al.*, 1999). Aujourd'hui, les termes synténie et colinéarité sont utilisés indifféremment pour décrire la conservation de l'ordre de groupes de gènes au sein d'une ou entre deux ou plusieurs espèces (Keller and Feuillet, 2000). La synténie est la conséquence et une preuve de l'origine évolutive commune des espèces.

Méthodologies de la génomique comparée

Dans les années 1960 les premières comparaisons entre génomes ont été réalisées (Chu and Swomley, 1961) en s'appuyant sur l'analyse de la taille et de la forme (position des centromères) des chromosomes par observation des caryotypes. Au fil du temps, la résolution de ce type d'approches, dites cytogénétiques, a progressé. Dans les années 1970, la méthode de *C-banding* (Hsu and Arrighi, 1971) a permis de différencier plusieurs régions au sein de chaque chromosome par coloration de la chromatine en fonction de son état de condensation sur les histones. Cette méthode a notamment permis de mettre en évidence la fusion de chromosomes

à l'origine de la divergence entre l'Homme et le chimpanzé (Turleau and Grouchy, 1973) et de détecter des réarrangements chromosomiques importants chez le blé (Gill and Kimber, 1974). Par la suite, les méthodes d'hybridation *in situ* ont augmenté la résolution des observations cytogénétiques, atteignant l'échelle du mégabase (10^6 paires de bases), ce qui a permis de caractériser finement les variations structurales au sein de génomes complexes tels que celui du blé tendre (Zhang *et al.*, 2004). Cependant, la cytogénétique présente la limitation majeure de ne pouvoir comparer les structures des génomes qu'entre espèces relativement proches évolutivement. Si deux espèces sont trop divergentes le signal d'homologie est brouillé par des différences de structures trop importantes (nombre de réarrangements, contenu en éléments transposables).

Pour proposer une vision exhaustive des liens entre l'ensemble des organismes vivants, la génomique comparative s'est appuyée sur des approches permettent de relier des espèces plus éloignées évolutivement. De ce point de vue, la comparaison des cartes génétiques a constitué un grand pas en avant. La cartographie génétique consiste à positionner des marqueurs liés à des caractères polymorphiques sur les chromosomes (ou groupes de liaisons) en fonction du taux de recombinaison entre les positions (loci) repérées par ces marqueurs entre individus d'une population en ségrégation. L'unité de mesure de la distance entre les marqueurs est le centimorgan (noté cM). La cartographie génétique s'est développée dès la première partie du 20^{ième} siècle avec la publication des premières cartes génétiques de drosophile dès 1913 (Sturtevant, 1913) puis du maïs en 1935 (Emerson *et al.*, 1935). Ces cartes se basaient sur des marqueurs morphologiques dont le nombre était limité par le nombre de caractères observables ce qui se traduisait par une très faible résolution. C'est dans les années 1980 que les cartes génétiques se sont développées grâce à l'avènement des marqueurs moléculaires. L'utilisation des marqueurs RLFP (*Restriction Fragment Length Polymorphism*) et SSR (*Simple Sequence Repeat*) a nettement amélioré la résolution des cartes génétiques, permettant à la fois la comparaison des chromosomes entre eux au sein d'un génotype, mettant notamment en évidence des duplications, et la comparaison de différentes cartes génétiques de différentes espèces ouvrant la voie à l'étude de la synténie. Ainsi la conservation de l'ordre des marqueurs entre espèces (synténie ou colinéarité) a été mise en évidence chez les monocotylédones, en

comparant les cartes du sorgho et du maïs (Whitkus et al., 1992, Figure 12), et chez les dicotylédones, en comparant les cartes de la tomate et de la pomme de terre (Gebhardt *et al.*, 1991).

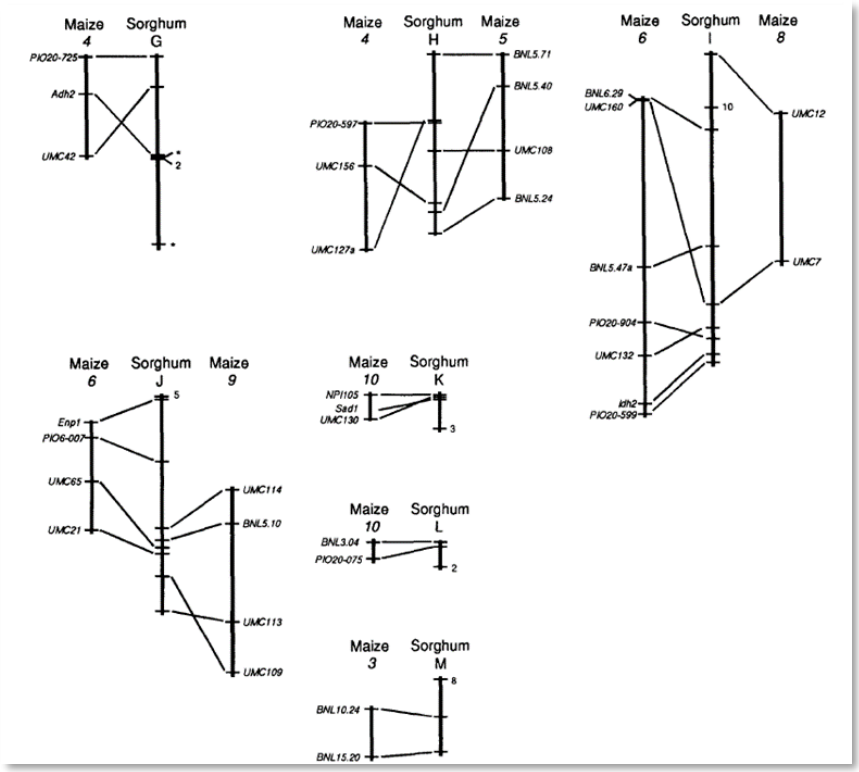


Figure 12 : comparaison des cartes génétiques de maïs et de sorgho sur la base de marqueurs RFLP. Les chromosomes sont matérialisés par des barres verticales et les marqueurs moléculaires conservés sont reliés par des traits.

Chez les céréales, la comparaison de cartes génétiques (Moore *et al.*, 1995) a permis la mise en évidence d'une forte colinéarité entre les marqueurs, permettant de modéliser le caryotype ancestral des céréales sur la base de la synténie et de larges blocs conservés entre ces espèces. En dépit d'une importante diversité de structures des génomes au sein de cette famille, notamment en termes de taille et de nombre de chromosomes, les comparaisons entre les cartes génétiques de plusieurs espèces de céréales ont montré une très forte conservation de la synténie autorisant à considérer cette famille, et au-delà l'ensemble des monocotylédones, comme un « système génétique unique » (Bennetzen and Freeling, 1993). Ces analyses bien que basées sur des comparaisons de cartes génétiques à faible résolution (un marqueur tous les 10 cM), montrent que, malgré une divergence de près de 60 millions d'années, moins de trente blocs de liaisons suffisent pour représenter tous les autres génomes de céréales en utilisant le génome

du riz en tant que pivot. Les différents blocs de liaison peuvent ainsi être utilisés à la manière de 'Lego' pour reconstruire les chromosomes de chaque espèce étudiée (Figure 13.A). De plus, les bornes de ces trente blocs coïncident fréquemment avec la position des centromères et télomères, ce qui suggère que ces sites jouent un rôle fondamental dans l'évolution des chromosomes des graminées (Moore *et al.*, 1997; Devos and Gale, 2000). C'est sur ces bases que Michael D. Gale, et son équipe au John Innes Centre, représentent la conservation de blocs de synténie entre les génomes de graminées issues des analyses de comparaisons de cartes génétiques par des cercles concentriques, les *Crop circles* (Moore *et al.*, 1995; Gale and Devos, 1998). Cette représentation permet d'identifier rapidement les chromosomes orthologues entre six espèces de graminées, à savoir le riz, le sorgho, le maïs, les Triticeae, le *foxtail* millet et la canne à sucre (Figure 13.B).

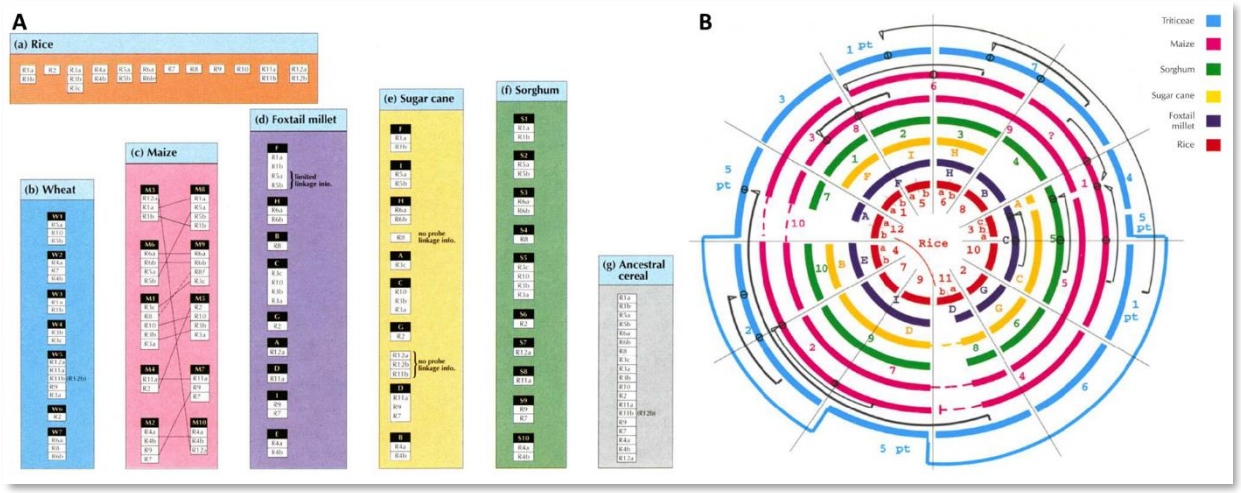


Figure 13 : A. Représentation sous forme de blocs des liens de synténie entre 6 espèces de céréales (a à f : Riz, Blé, Maïs, Millet, Canne à sucre, Sorgho) et (g) reconstruction de la structure de leur ancêtre commun. B. Représentation de ces mêmes liens de sous forme de cercles concentriques Crop circles. Figure adaptée de Moore et al. 1995.

Le développement de nouveaux marqueurs (notamment SNP : *Single Nucleotide Polymorphism*) et les nouvelles technologies de génotypage par séquençage ont permis d'enrichir significativement les cartes génétiques afin de réaliser des comparaisons beaucoup plus résolutives notamment chez les céréales (Salse *et al.*, 2008). Cependant, ces analyses de génomique comparative utilisant les données des cartes génétiques, à l'instar des approches cytogénétiques, sont limitées par la résolution de la méthode de détection et elles ne permettent que de détecter les plus grands réarrangements. De plus, la construction de cartes génétiques

peut s'avérer complexe et fastidieuse pour certaines espèces notamment en fonction de la taille de la population nécessaire pour générer une haute résolution de marquage, mais aussi de la possibilité d'avoir des parents contrastés augmentant le taux de polymorphisme entre marqueurs. Face à ces limitations, l'entrée dans l'ère du séquençage des génomes végétaux depuis le début des années 2000 a ouvert de nouvelles perspectives pour caractériser et comparer les génomes (Soltis *et al.*, 2013).

2.2 La paléogénomique : reconstruire les génomes pour retracer l'histoire évolutive

2.2.1 Reconstruire les génomes disparus pour révéler les mécanismes évolutifs

La paléogénomique a pour objet de reconstruire la structure de génomes ancestraux à l'origine des espèces modernes, afin de comprendre leurs histoires évolutives. Le champ de la paléogénomique se divise en deux types d'approches mises en œuvre pour reconstruire le génome ancestral. La première méthode consiste à retrouver puis à séquencer de l'ADN fossile. Elle a permis de déchiffrer les génomes de mammifères tels que le mammouth (Poinar *et al.*, 2006) et l'homme de Neandertal (Green *et al.*, 2010), puis de plantes, le coton (Palmer *et al.*, 2012) ouvrant la voie à des nombreuses espèces d'angiospermes (Di Donato *et al.*, 2018). La seconde approche, dite *in silico*, consiste à comparer les séquences des génomes d'espèces actuelles pour identifier les gènes conservés entre les espèces modernes, qui permettent de modéliser le génome ancestral commun défini comme le génome minimal théorique dans sa structure chromosomique et son contenu en gènes. Un génome ancestral est donc un génome "médian" ou "intermédiaire" constitué d'un ensemble de groupes de gènes conservés entre espèces modernes et considérés comme les protochromosomes ancestraux. Ainsi, les premiers génomes de mammifères séquencés ont permis de reconstruire le caryotype ancestral commun à l'Homme, à la souris et au rat (Bourque *et al.*, 2004) puis, en fonction de la disponibilité de nouvelles séquences de génome, l'ensemble de l'évolution des mammifères placentaires a pu être retracé (Nakatani *et al.*, 2007).

Il est notable que les deux approches sont complémentaires en termes d'échelles de temps et d'objets étudiés : temps court à moyen (d'une centaine à quelques dizaines de milliers d'années) et étude des dynamiques populationnelles et de domestication pour la stratégie de séquençage

des ADN anciens ; temps long (jusqu'à des centaines de millions d'années) et étude des dynamiques de spéciation pour la stratégie de reconstruction *in silico* (Pont, *et al.*, 2019).

2.2.2 Principe de la reconstruction 'in silico' des génomes ancestraux

La liste et l'ordre des gènes constituant un protochromosome sont déduits par comparaison du contenu en gènes des espèces modernes grâce à une stratégie en quatre étapes (Figure 14)(Pont, *et al.*, 2019).

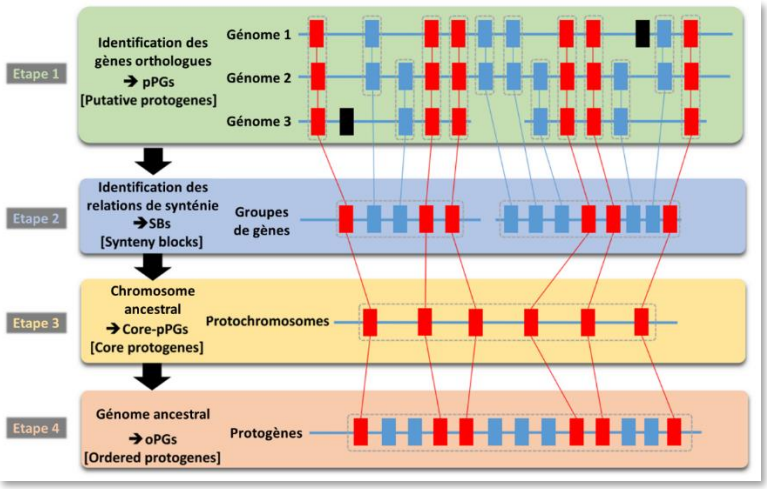


Figure 14 : Méthode de reconstruction des génomes ancestraux. Étape 1 : identification des gènes conservés, dits protogènes putatifs (pPG). Les gènes sont figurés par des rectangles et reliés par des lignes rouges lorsqu'ils sont conservés entre toutes les espèces étudiées et par des lignes bleues lorsqu'ils sont conservés dans un sous-ensemble de ces espèces. Les gènes spécifiques à une seule espèce sont représentés par des rectangles noirs. Étape 2 : identification de la synténie entre groupes de gènes adjacents conservés pour constituer des blocs de synténie (SBs), qui sont mis en évidence par des rectangles gris en pointillé. Étape 3 : reconstruction des régions ancestrales contiguës (CARs) contenant les gènes conservés dans toutes les espèces étudiées (référencées comme core-pPGs). Étape 4 : reconstruction du génome ancestral fournissant des protochromosomes intégrant les protogènes ordonnés (oPGs).

La première étape consiste à comparer les séquences annotées de l'ensemble des génomes considérés pour identifier les paires de gènes conservés ou dupliqués sur la base des paramètres d'alignement stricts (Salse *et al.*, 2009) pour définir les gènes conservés entre paires d'espèces. Les gènes ainsi identifiés constituent des protogènes putatifs (pPG). La deuxième étape consiste à identifier les pPG qui sont conservés entre toutes les espèces étudiées. Notés core-pPGs, ils définissent des blocs de synténie, *synteny blocks* (SB). La troisième étape consiste à fusionner les SB sur la base des relations orthologues chromosome à chromosome entre les génomes comparés pour définir des protochromosomes ancestraux, également appelés régions ancestrales contiguës (*contiguous ancestral region*, CARs). Les CARs représentent des ensembles

indépendants de blocs génomiques présentant des relations paralogues et/ou orthologues entre les espèces modernes. La quatrième et dernière étape consiste à enrichir les contenus en gène des CARs en intégrant l'ensemble des pPGs, y compris ceux qui ne sont conservés que dans un sous-ensemble d'espèces pour obtenir l'ensemble exhaustif des protogènes ordonnés (oPG). Les gènes orthologues putatifs qui ont été transposés en dehors des CARs, non-conservés en synténie au cours de l'évolution ne sont pas identifiés dans les SB et sont donc absents des génomes ancestraux reconstruits.

Les caryotypes ancestraux ainsi reconstruits vont permettre de proposer un scenario évolutif sur la base du chemin le plus parcimonieux, c'est-à-dire introduisant le plus petit nombre de réarrangements génomiques (inversions, délétions, translocations, fusions, fissions, duplications) pour expliquer la transition entre les génomes ancêtres et modernes (Salse *et al.*, 2008). Un tel modèle évolutif permet d'étudier les trajectoires évolutives à l'échelle des génomes ou des familles de gènes spécifiques grâce à l'identification des changements strucutraux impliqués et leur affectation à des espèces ou des familles botaniques particulières ou à des caractères spécifiques.

2.2.3 Enseignements de la reconstruction des génomes ancestraux

En 2019, Pont et al. présentent une reconstruction complète de l'histoire évolutive des angiospermes (Pont, *et al.*, 2019). Les angiospermes ont évolué depuis 190 à 250 millions d'années, à partir d'un ancêtre commun à 15 protochromosomes constituants le caryotype ancestral des angiospermes, *ancestral angiosperms karyotype* (AAK). Cette période correspond à la fin de l'ère triasique et précède le premier fossile végétal retrouvé (Herendeen *et al.*, 2017). L'AAK a ensuite divergé en 2 branches pour donner l'ancêtre des monocotylédones, *ancestral monocot karyotype* (AMK), présentant cinq protochromosomes, et celui des dicotylédones, *ancestral eudicot karyotype* (AEK), avec sept protochromosomes. Il est possible de reconstruire le caryotype de n'importe quel génome d'angiosperme moderne sous forme de mosaïque de segments de protochromosomes, soit à partir de AAK ou, si l'on souhaite étudier spécifiquement l'histoire évolutive des monocotylédones ou des dicotylédones, à partir de partir d'AMK ou d'AEK respectivement. Cette mosaïque est représentée par le *painting* des chromosomes modernes en fonction des couleurs attribuées aux protochromosomes ancestraux (Figure 15).

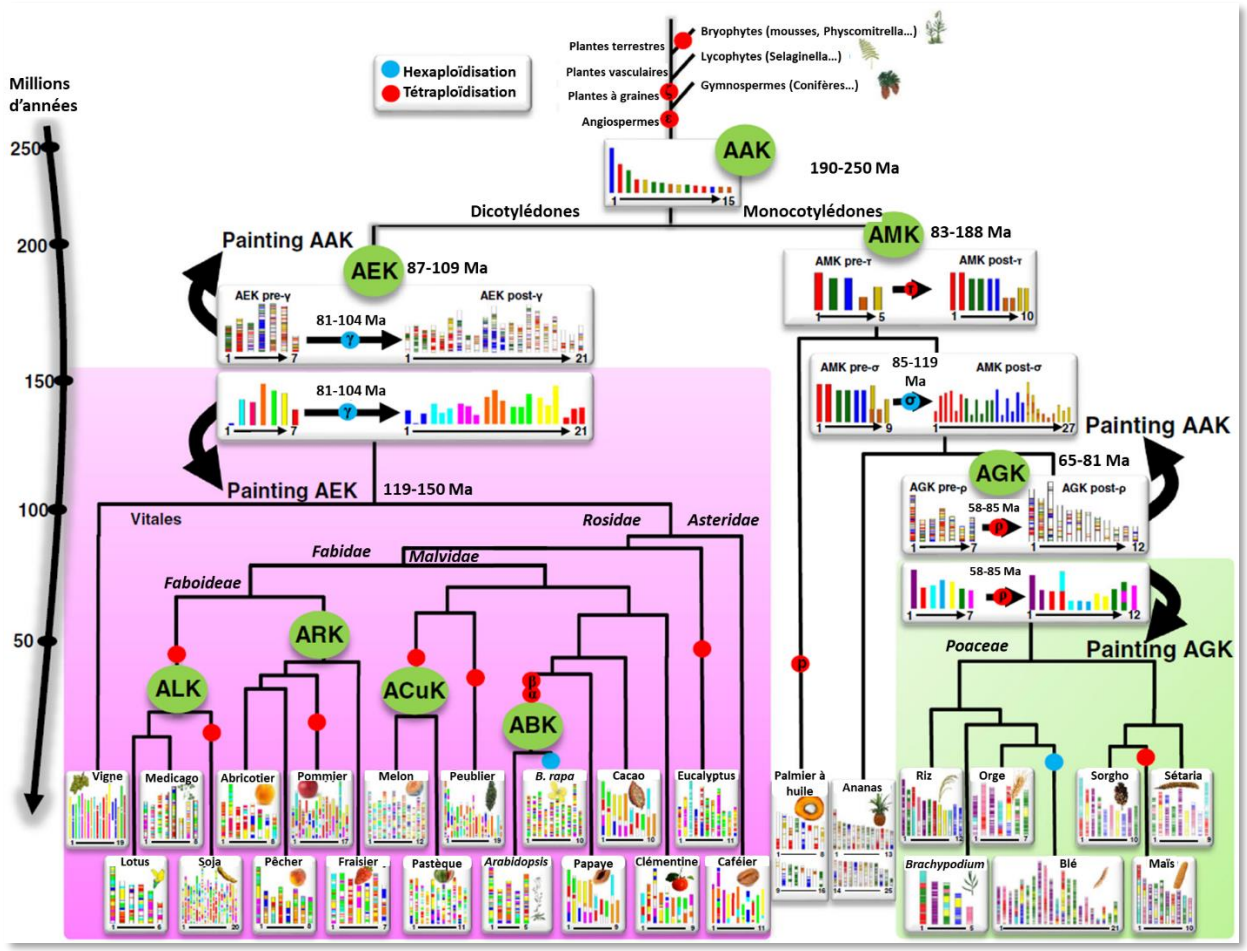


Figure 15 : Analyse de l'évolution des génomes des angiospermes sur la base la reconstruction des caryotypes ancestraux. Les génomes actuels des monocotylédones (à droite, avec les graminées sur fond vert) et des dicotylédones (à gauche, sur fond rose) sont représentés avec des codes de couleur retraçant l'évolution des segments génomiques à partir de leurs ancêtres fondateurs. La datation des évènements figurés est représentée par l'échelle de temps indiquée à gauche (en Ma). Les paintings AGK et AEK (en bas) représentent les réarrangements des caryotypes des génomes modernes des monocotylédones et des dicotylédones, par rapport aux caryotypes ancestraux, respectivement, des graminées (AGK) et des dicotylédones (AEK), présentant tous deux sept protochromosomes (code couleur). Sur le même principe, le painting AAK (en haut) représente AGK, AEK et le caryotype monocotylédone ancestral (AMK), ainsi que les génomes du palmier à huile et de l'ananas, sur la base du caryotype angiosperme ancestral (AAK) de 15 protochromosomes (code couleur). Les événements de polyploïdisation qui ont façonné la structure des génomes des plantes modernes par WGD des génomes ancestraux déduits sont indiqués par des points rouges (duplication / tétraploïdisation) et des points bleus (triplication / hexaploïdisation). Adapté de Pont, et al. 2019.

Au-delà de la structure des caryotypes, les reconstructions des génomes ancestraux permettent l'étude de l'évolution fonctionnelle des génomes en fonction des enrichissements ou des déplétions des gènes annotés pour des fonctions particulières, organisées en différentes ontologies de termes (*gene ontology* (GO). A titre d'exemple, la comparaison du répertoire des gènes de AAK à celui d'espèces évolutivement proches des angiospermes mais ne présentant pas

certains caractères particuliers (fleurs, vascularisation, différenciation des organes), telles que les gymnospermes, les mousses et les algues vertes unicellulaires, a permis de découvrir des gènes spécifiques aux plantes à fleurs. Les GO de ces gènes sont associés à des processus tels que l'interaction entre pollen et pistil, la réponse aux stimuli endogènes, le développement de la fleur et la pollinisation, autant de processus biologiques clés pour la transition entre les gymnospermes et les angiospermes (Murat *et al.*, 2017).

Outre la construction des caryotypes des 3 ancêtres les plus anciens AMK, AEK et AAK, l'approche de paléogénomique par comparaison *in silico* a permis de reconstruire des génomes ancestraux de référence pour les principales lignées d'angiospermes. Chez les dicotylédones, des génomes ancestraux ont été proposés pour les sous-familles des *Rosaceae* (Raymond *et al.*, 2018), des *Brassicaceae* (Murat, *et al.*, 2015), et *Cucurbitaceae* (Wu *et al.*, 2017), présentant respectivement en 9, 8 et 12 protochromosomes. L'ancêtre des légumineuses a également été analysé (J., Wang *et al.*, 2017). Parmi les monocotylédones, le caryotype ancestral de la famille des graminées (AGK), qui inclut le riz, le blé, l'orge, *Brachypodium*, le sorgho, la sétaria et le maïs, présente sept protochromosomes (Murat *et al.*, 2014) (Tableau 3).

Famille	Date (Ma)	Acronyme	# protochromosomes	# protogènes	Reference
Angiospermes	190–250	AAK (post-ε/ζ)	15	22 899	Murat *et al.* , 2017
Dicotylédones	87–109	AEK (pre-γ)	7	6 284	Murat *et al.* , 2017
Dicotylédones	87–109	AEK (post-γ)	21	9 022	Murat *et al.* , 2017
Monocotylédones	100–150	AMK (pre-τ)	5	6 707	Murat *et al.* , 2017
Monocotylédones	100–150	AMK (post-τ)	10	13 916	Murat *et al.* , 2017
Graminées	65–81	AGK (pre-ρ)	7	8 581	Murat *et al.* , 2014
Graminées	70–96	AGK (pre-ρ)	7	9 430	Wang *et al.* , 2015
Graminées	65–81	AGK (post-ρ)	12	16 464	Murat *et al.* , 2014
Graminées	70–96	AGK (post-ρ)	12	18 86	Wang *et al.* , 2015
Brassicaceae	27–40	ABK (post-α/β)	8	20 037	Murat *et al.* , 2015
Brassicaceae	23–27	ACaK (post-α/β)	8	22 085	Murat *et al.* , 2015
Brassicaceae	23–27	PCK (post-α/β)	7	21 227	Murat *et al.* , 2015
Rosaceae	70–90	ARK (post-WGD)	9	8 861	Raymond *et al.* , 2018
Cucurbitaceae	25–50	ACuK (post-WGD)	12	18 534	Wu *et al.* , 2017
Légumineuses	56–59	ALK (post-WGD)	–	28 900	Wang *et al.* , 2017

Tableau 3 : Liste des reconstructions de caryotypes ancestraux de génomes d'angiospermes, avec la famille botanique ciblée, la date (en million d'années Ma), l'acronyme du génome ancestral, le statut par rapport aux évènements de polyploïdisation, le nombre de protochromosomes, le nombre de protogènes, les références dans la littérature. Abréviations : AAK Ancestral Angiosperm Karyotype, ABK Ancestral Brassicaceae Karyotype, ACaK Ancestral Camelineae Karyotype, ACuK Ancestral Cucurbitaceae Karyotype, AEK Ancestral Eudicot Karyotype, AGK Ancestral Grass Karyotype, ALK Ancestral Legume Karyotype, AMK Ancestral Monocot Karyotype, ARK Ancestral Rosaceae Karyotype, PCK Proto-Calepineae Karyotype, WGD Whole Genome Duplication.

L'analyse de l'histoire évolutive des graminées met en évidence un évènement initial de duplication complète du génome, aussi désignée par l'acronyme WGD pour *Whole Genome Duplication*, il y a plus de 95 millions d'années. Cet évènement subit par les graminées et détecté grâce à la comparaison des contenus en gènes des espèces actuelles avec celui du caryotype ancestral est typique du passage à l'état polyploïde qui est identifié dans toutes les branches de l'arbre évolutif des angiospermes. La description du phénomène de polyploïdie est l'objet de la section suivante.

3 La Polyploïdie, évènement ubiquitaire et moteur de l'évolution

Chez les plantes, la polyploïdie est un phénomène répandu et aux conséquences majeures sur l'évolution des génomes. Dans cette section, la polyploïdie sera définie précisément, puis les mécanismes qui aboutissent à son apparition, ses conséquences à de multiples échelles sur la biologie de la plante, ses impacts sur la structure et la régulation du génome seront présentés

3.1 Définitions et nomenclature

Un organisme est dit diploïde lorsque chacune de ses cellules contient deux copies de chaque chromosome formant une paire, alors qu'il sera dit polyploïde si chacune de ses cellules contient plus de deux copies de chaque chromosome. L'Homme est un organisme diploïde, mais de nombreux êtres vivants sont des polyploïdes. Le phénomène de polyploïdie a d'abord été observé chez les plantes par les cytogénéticiens dès les années 1920 (McClintock, 1929). En 1970, Susumu Ohno développe l'hypothèse que la duplication des gènes, conséquence de la polyploïdie, joue un rôle majeur dans l'évolution des espèces (Susumu Ohno, 1970). La polyploïdie est dès lors un sujet d'étude notamment dans les années 1990 (Sidow, 1996; Meyer and Schartl, 1999). Mais ça n'est qu'au début des années 2000, en s'appuyant sur l'avènement du séquençage des génomes, que le caractère ubiquitaire de la polyploïdie chez les eucaryotes a été révélé, notamment grâce à des travaux portant sur les champignons (Kellis *et al.*, 2004), les vertébrés (Dehal and Boore, 2005), les plantes angiospermes (Thomas *et al.*, 2006) et les ciliés (Aury *et al.*, 2006).

Parmi les eucaryotes, le règne des *Plantae* est celui où le phénomène de polyploïdie est le plus fréquent (Wolfe, 2001). En effet, toutes les espèces végétales ont expérimenté au moins un événement de passage à l'état polyploïde, ou polyploïdisation, au cours de leurs histoires évolutives (Van De Peer *et al.*, 2017). Ainsi, il a été montré chez les plantes à fleurs, que les eudicotylédones présentent une triplication ancestrale du génome (référencée γ, Jaillon *et al.*, 2007) tandis que toutes les monocotylédones partagent une duplication ancestrale (référencée τ, Jiao *et al.*, 2014). Au sein des monocotylédones, les céréales ont subi ensuite deux nouvelles duplications de leur génome (référencée σ et ρ, Jiao *et al.*, 2014). Toutes les espèces angiospermes actuelles sont donc des paléopolyploïdes, notamment, la plante modèle Arabidopsis (Thomas *et al.*, 2006) et les principales cultures comme le Sorgho (Paterson *et al.*, 2009), le maïs (Schnable *et al.*, 2011), les brassicacées (Cheng *et al.*, 2012; Murat, Louis, *et al.*, 2015), le blé (Pont *et al.*, 2013), le coton (Renny-Byfield *et al.*, 2015; Paterson *et al.*, 2012) et le soja (Schmutz *et al.*, 2010).

Pour indiquer la structure chromosomique d'un génome, on note *n* le nombre de chromosomes uniques et *x* le nombre de chromosomes dans le set de base. *n* correspond au nombre de chromosomes présent dans une cellule haploïde telle qu'un gamète et *2n* correspond au nombre total de chromosomes dans une cellule somatique. Dans le cas d'un organisme diploïde *n=x* et le nombre total de chromosomes dans une cellule somatique correspond à deux fois le set de base *2n=2x*. Dans le cas d'un organisme tétraploïde le nombre de total de chromosomes correspond à 4 fois le set de base soit *2n=4x*. Selon cette notation le nombre de chromosomes d'*Arabidopsis thaliana*, diploïde, est noté *2n=2x=10*, celui du blé dur, tétraploïde, *2n=4x=28* (Figure 16), et celui de la patate douce, hexaploïde, *2n=6x=90*.

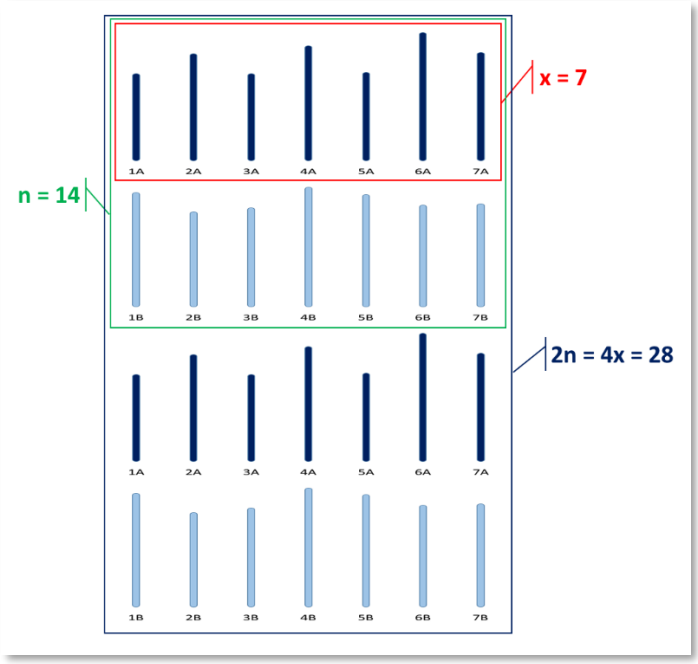


Figure 16 : Illustration de la nomenclature décrivant la structure chromosomique d'un génome. Dans cet exemple le blé tétraploïde présenté possède un set de base de x=7 chromosomes, n=14 chromosomes uniques dont 7 issus d'un progéniteur et 7 de l'autre, et un total de 28 chromosomes dans les cellules somatiques soit 2n=4x.

Les polyploïdes sont classiquement divisés en deux catégories en fonction de leurs origines. Les autopolypoïdes sont le produit d'un doublement chromosomique au sein d'une espèce, tandis que les allopolyploïdes résultent de l'hybridation des génomes de deux espèces apparentées (Stebbins, 1947). Cependant, cette dichotomie est à considérer avec précaution. Le degré de divergence entre les deux espèces dans le cas de l'allopolyploïdie est très variable. Ainsi, certains événements d'hybridations interspécifiques, qui conduisent à une allopolyploïdie, peuvent impliquer des espèces dont les sous-génomes sont moins divergents les uns des autres que les événements d'hybridation au sein d'une seule espèce hautement polymorphe (Mason and Wendel, 2020). Il est donc raisonnable de considérer qu'il y a un continuum entre allopolyploïdie et autopolyploïdie. Lorsque l'évènement étudié est ancien et que des modifications successives du génome du néopolyploïde se seront combinées à celles intervenues suite à l'évènement initial, il est difficile de définir la nature de l'évènement de duplication initial.

3.2 Mécanismes à l'origine de la polyploïdie

Quel que soit le type de polyploïdie, l'évènement initial qui induit le doublement chromosomique est soit un doublement somatique, soit la fusion de gamètes non réduits. Le passage de l'état diploïde à l'état tétraploïde se fait ensuite selon différentes modalités, soit directement, soit en passant par un état triploïde (Figure 17).

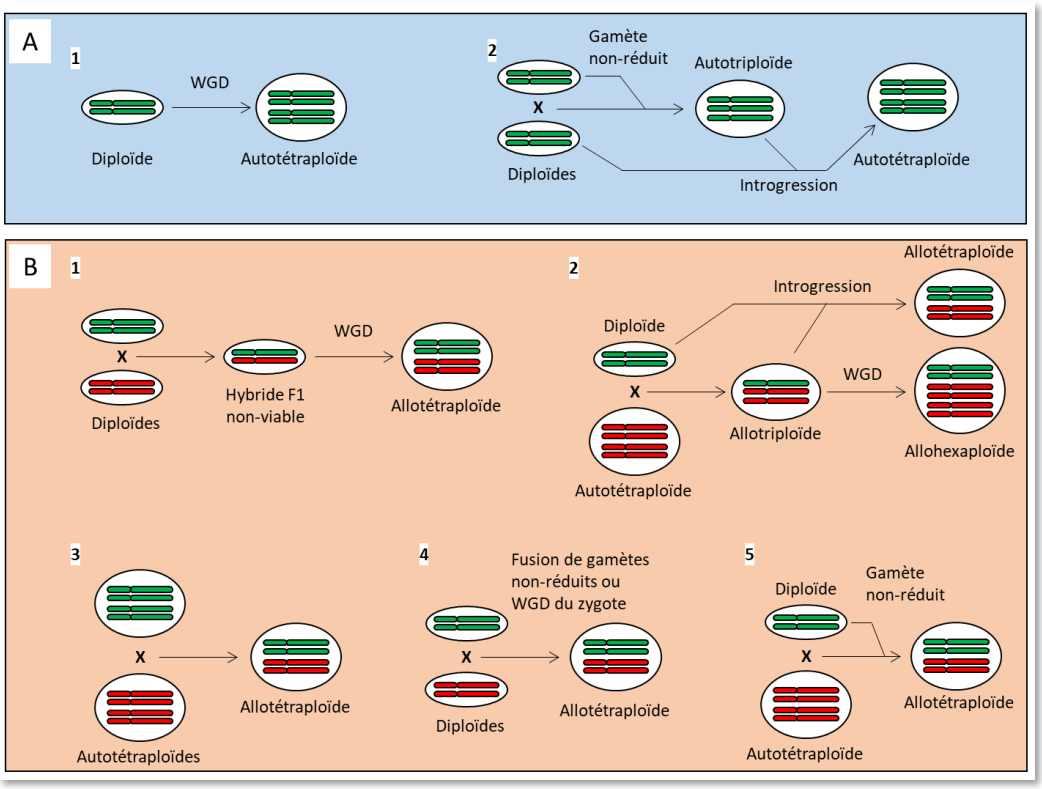


Figure 17 : Modes de formation des polyploïdes. A. Cas des génomes autopolyploïdes. A.1. Des gamètes non réduits peuvent fusionner pour donner immédiatement naissance à une descendance autotétraploïde. A.2. Les gamètes non réduits peuvent se combiner avec des gamètes haploïdes normaux pour donner des triploïdes. Ceux-ci peuvent alors produire des gamètes aneuploïdes et, à un faible taux, des gamètes possédant le nombre attendu de chromosomes qui ont le potentiel de donner une descendance qui, avec le temps, se stabilisera en génotypes entièrement diploïdes et entièrement tétraploïdes. Cela peut également faciliter le flux génétique entre les ploïdies via un "pont triploïde". B. Cas des génomes allopolyploïdes. B.1. Un hybride stérile se forme par la fusion de gamètes haploïdes normaux issus de deux espèces diploïdes distinctes, puis devient fertile grâce à un évènement de WGD qui offre aux chromosomes des partenaires d'appariement. B.2. Un hybride stérile se forme par la fusion de gamètes haploïdes normaux issus de deux espèces, l'une diploïde, l'autre tétraploïde. Il devient fertile sous la forme, soit d'un allotétraploïde en constituant un pont triploïde, soit d'un allohexaploïde grâce à un évènement de WGD. B.3. La fusion des gamètes euploïdes issus de deux espèces autotétraploïdes distinctes produit une espèce allotétraploïde. B.4. La fusion de deux gamètes non-réduits issus de deux espèces distinctes ou la WGD d'un zygote euploïde donnent immédiatement naissance à une descendance allotétraploïde. B.5. La fusion entre le gamète non réduit d'une espèce diploïde et le gamète euploïde d'une espèce autotétraploïde donne naissance à une espèce allotétraploïde.

La fréquence de la production de gamètes non-réduits augmentent lorsque les plantes se trouvent dans des conditions de stress (Mason and Pires, 2015), telles que des températures extrêmes ou des périodes de sécheresse (Pécrix *et al.*, 2011; de Storme *et al.*, 2012). De telles conditions auront potentiellement un impact sur le taux de polyploïdie des espèces qu'elles affectent. Ce phénomène est largement étudié chez la levure, organisme eucaryote unicellulaire possédant un set de base de 16 chromosomes, viable à divers niveaux de ploïdie : haploïde, diploïde et tétraploïde. Des expériences de culture en conditions contrôlées portant sur plusieurs dizaines de générations démontrent que des conditions de stress entraînent la variation du niveau de ploïdie pour les souches haploïdes et un maintien de l'état polyploïde pour les souches tétraploïdes. C'est notamment le cas pour des milieux carencés en azote (Gerstein *et al.*, 2017), présentant un stress salin (Gerstein *et al.*, 2006) ou soumis à une augmentation de la température (Yona *et al.*, 2012).

3.3 Effets de la WGD sur la biologie de la plante

La polyploïdisation a été proposée comme un moteur majeur de l'évolution des plantes fournissant un nouveau matériel génétique (Freeling, 2017; Van de Peer *et al.*, 2021). Elle a des conséquences notamment sur le développement et la reproduction des plantes. Ces conséquences décrites ci-après se traduisent à différents niveaux de la biologie de la plante.

3.3.1 Impacts de la WGD sur la méiose

Les nouveaux organismes polyploïdes doivent rapidement mettre en place une différenciation des sous-génomes pour assurer une méiose fonctionnelle afin de produire des gamètes équilibrés qui permettront, après fécondation, de restaurer le niveau de ploïdie parental et, ainsi, de générer une descendance fertile (Lloyd and Bomblies, 2016). Chez les diploïdes, lors de la méiose les chromosomes homologues s'associent en bivalents, puis ségrégent lors de la première division, puis les chromatides sœurs se séparent lors de la seconde division constituant le matériel génétique des gamètes (*n*) équivalant à la moitié du set de chromosomes parental. Le niveau de ploïdie parental (*2n*) sera restauré par la fécondation.

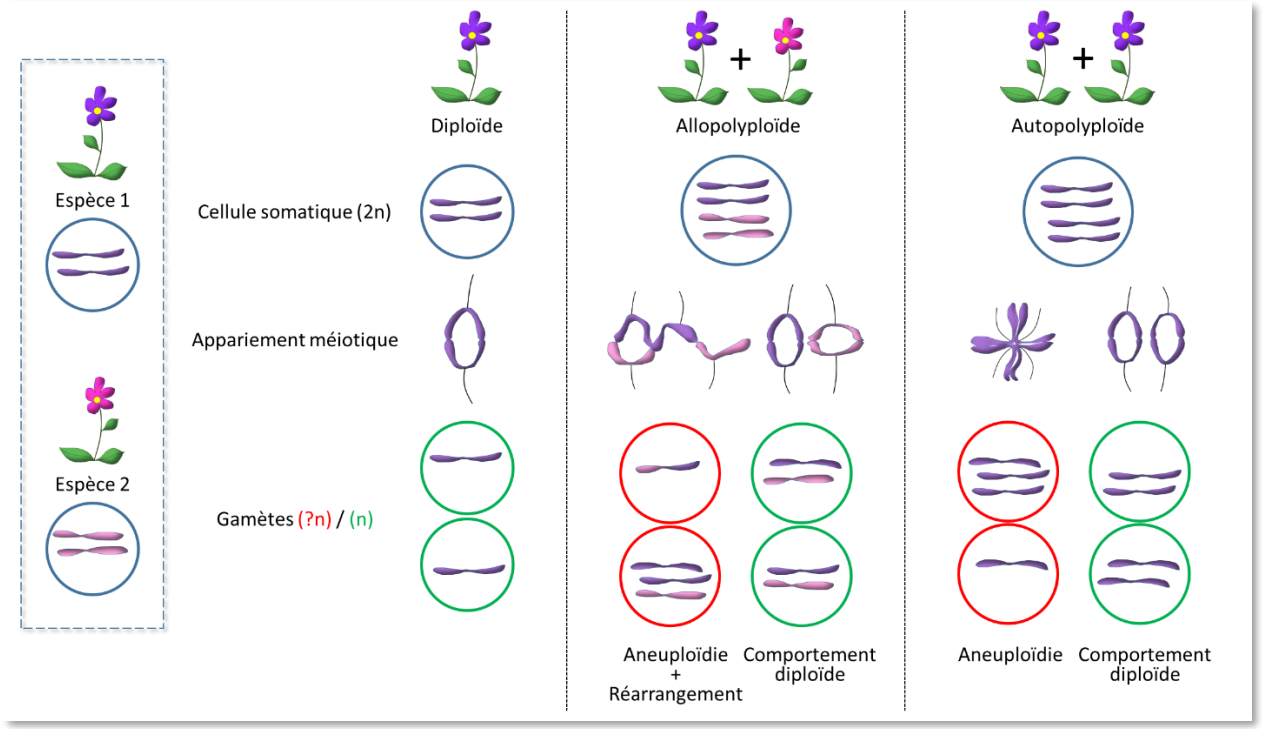


Figure 18 : Les deux génomes diploïdes constitutifs des allopolyploïdes sont représentés en rose et violet. Pour les allopolyploïdes comme pour les autopolyploïdes, les méioses aberrantes sont représentées à gauche et les appariements fonctionnels à droite. Les allopolyploïdes nouvellement évolués présentent souvent un appariement aberrant entre les homéologues pendant la méiose, pouvant entraîner des réarrangements chromosomiques, la perte de gènes et l'aneuploïdie. Les autopolyploïdes peuvent présenter un appariement multivalent entre les quatre chromosomes homologues qui sera problématique pour la ségrégation des chromosomes dans les gamètes, conduisant à l'aneuploïdie. En général, les allopolyploïdes et les autopolyploïdes fertiles ont rétabli un comportement d'appariement chromosomique de type diploïde. Adapté de Hollister, 2015.

La WGD constitue un défi pour la ségrégation chromosomique lors de la méiose (Figure 18) car elle offre à chaque chromosome plus d'une association possible (Hollister, 2015). Or, l'implication d'un chromosome dans une association illégitime ou multiple, conduit à la production de gamètes aneuploïdes. En conditions stables celles-ci présentent une fertilité amoindrie et la probabilité réduite de donner naissance à des descendants eux-mêmes fertiles, ce qui impacte le succès reproducteur, ou *fitness*, du génotype concerné (Mercier *et al.*, 2015). Alors que de nombreux néopolyploïdes présentent des anomalies méiotiques dont l'ampleur et le type diffèrent en fonction de l'origine de la polyploïdie, les polyploïdes établis présentent généralement un comportement d'appariement de type diploïde, c'est-à-dire un appariement bivalent (Comai, 2005).

Dans le cas des autopolyploïdes toutes les copies de chaque chromosome sont véritablement homologues et ont la même chance de se recombiner entre elles. La conséquence est la formation fréquente de multivalents chiasmatiques, quoique non systématique, observée chez les autopolyploïdes qu'ils soient naturels ou artificiels (Grandont *et al.*, 2013). Cependant des mécanismes peuvent se mettre en place pour parvenir à une méiose équilibrée. Le plus connu est la réduction du nombre de chiasmata, ou crossing-over, les points de contacts entre les chromatides de chromosomes homologues, ce qui augmente la spécificité des appariements. Ce phénomène a notamment été montré chez l'autotétraploïde *Arabidopsis arenosa*, qui présente moins de chiasmata par bivalent que les espèces apparentées diploïdes (Yant *et al.*, 2013).

Chez les allopolyploïdes, qui ont une origine hybride interspécifique, la situation est différente car les chromosomes qui proviennent de différents parents diploïdes (c'est-à-dire les homéologues) ne sont pas totalement interchangeables. Lors de la première phase de la méiose le strict appariement spécifique des chromosomes homologues entre eux, et son corollaire, l'absence d'appariement avec les chromosomes homéologues, sont nécessaires pour assurer une ségrégation chromosomique correcte. Ce processus est contrôlé génétiquement chez la plupart, sinon la totalité, des espèces allopolyploïdes (Jenczewski and Alix, 2004), mais malgré leur rôle crucial, notre connaissance des loci régulant la formation de *crossing overs* entre les chromosomes homéologues demeure limitée. Le locus le mieux caractérisé est le locus Ph1 du blé cartographié sur le chromosome 5B (Riley and Chapman, 1967; Dubcovsky *et al.*, 1995) qui favorise une méiose de type diploïde par des altérations de la condensation de la chromatine.

3.3.2 Augmentation de la taille des cellules et effets induits

Il a été montré que la taille, et donc le volume, d'une cellule augmentent en fonction de la quantité d'ADN contenue dans le noyau (D'Ario and Sablowski, 2019) tant chez les vertébrés (Olmo, 1983) que chez les plantes (Price, 1988). C'est logiquement le cas lorsque l'augmentation de la quantité d'ADN est due à une polyploïdisation (Doyle and Coate, 2019).

Cette augmentation de la taille des cellules affecte tous les types cellulaires. Elle provoque par conséquent un effet 'gigantisme' à l'échelle de la plante entière induisant l'augmentation de la taille des fleurs, des feuilles ou des fruits, autant de caractères d'intérêt agronomiques (Rosellini *et al.*, 2016; Rho *et al.*, 2012; Ruiz *et al.*, 2020). Au-delà des bénéfices directs de l'augmentation de la taille des organes, la taille des cellules est susceptible d'influer sur plusieurs mécanismes physiologiques et développementaux fondamentaux de la plante. L'augmentation de la taille des stomates entraine une augmentation des échanges gazeux de nature à expliquer la stimulation de la photosynthèse observée chez de nombreuses plantes polyploïdes (Bomblies, 2020; Drake *et al.*, 2013).

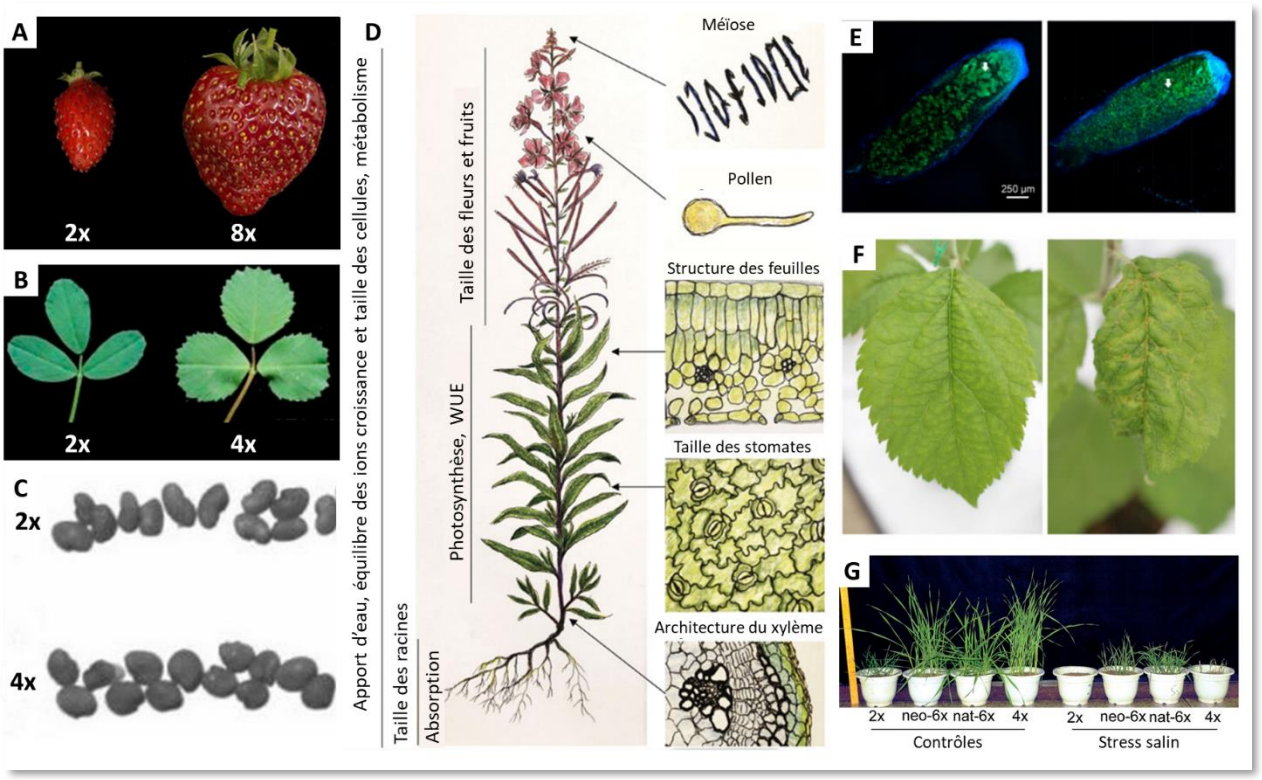


Figure 19 : Impacts de la polyploïdie. A. B. et C. Effets gigantisme sur les fruits du fraisier (Li et al. 2021), et sur les feuilles et les graines de Medicago sativa (Rosellini et al. 2016). D. Présentation des effets sur divers processus physiologiques à l'échelle de différents organes et tissus de la plante (Bomblies, 2020). E. Effets sur la symbiose : des nodules racinaires de luzerne néotétraploïde (à gauche) et diploïde synthétique (à droite) inoculés avec des rhizobia. Les plantes autotétraploïdes produisent des nodules 20 % plus grands que les diploïdes (Forrester et Ashman, 2020). F. Réponse de pommiers tétraploïdes (à gauche) et diploïdes (à droite) à l'infection fongique causant la maladie de la tavelure du pommier. Les symptômes visuels et la quantification par PCR de l'agent pathogène V. inaequalis diminuent dans le tétraploïde par rapport à son progéniteur diploïde (Hias et al., 2018). G. Résistance au stress Salin améliorée chez les blés polyploïdes par rapport à leur progéniteur diploïde (yang et al, 2014).

L'augmentation de la taille des cellules peut aussi avoir des conséquences spécifiques dans les interactions biotiques comme la relation des légumineuses avec leurs partenaires symbiotiques fixateurs d'azote. Ainsi, des lignées tétraploïdes de luzerne, *Medicago sativa*, tirent un meilleur parti du mutualisme avec les bactéries symbiotiques *Sinorhizobium* que les lignées diploïdes (Forrester *et al.*, 2020). Ces lignées tétraploïdes présentent de plus grands nodules avec des zones de fixation de l'azote plus étendues, permettant une fixation de l'azote atmosphérique plus importante. Cet avantage pourrait provenir de multiples effets, dont la taille des cellules. D'autres traits tels que la structure des feuilles, l'architecture du xylème ou le développement racinaire sont également modifiés par le changement de l'augmentation de taille des cellules (Bomblies, 2020) (Figure 19).

3.3.3 Résistances des polyploïdes aux stress biotiques et abiotiques

Outre ces changements induits par l'augmentation de la taille des cellules, la polyploïdie confère des avantages spécifiques, notamment des résistances aux stress abiotiques et biotiques (Te Beest *et al.*, 2012).

Stress abiotiques

De nombreuses études démontrent la capacité des plantes polyploïdes à résister à des stress biotiques tels que des températures extrêmes (froid, chaleur), des déficits hydriques ponctuels ou récurrents, et des milieux pauvres en nutriments ou potentiellement toxiques tels que des milieux salins (Tableau 4).

Espèce	Niveau de ploïdie	Tolérance	Référence
Arabidopsis *(Arabidopsis thaliana)*	Autotétraploïde	Carence en bore	Kasajima *et al.* , 2010
Arabidopsis *(Arabidopsis thaliana)*	Autotétraploïde	Sécheresse, stress salin	del Pozo and Ramirez-Parra, 2014
Arabidopsis *(Arabidopsis thaliana)*	Autotétraploïde	Stress salin	Chao *et al.* , 2013
Brachypodium *(B. distachyon)*	Autotétraploïde	Sécheresse (aridité)	Manzaneda *et al.* , 2012
Navet *(Brassica rapa L.)*	Autotétraploïde	Stress salin	Meng *et al.* , 2011
Cenchrus *(Cenchrus species)*	Allotétraploïde/allohexaploïde	Sécheresse	Chandra and Dubey, 2010
Centaurée du Rhin *(Centaurea stoebe)*	Autotétraploïde	Sécheresse	Mraz *et al.* , 2014
Lime *(Citrus limonia)*	Autotétraploïde (porte-greffe)	Sécheresse, froid, déficit en nutriments	Allario *et al.* , 2013
Lime *(Citrus limonia)*	Autotétraploïde (porte-greffe)	Stress salin combiné à la sécheresse	Allario *et al.* , 2011
Chrysanthème *(D.nankingense)*	Autotétraploïde	Sécheresse, stress salin, froid	Liu *et al.* , 2011
Igname *(Dioscorea zingiberensi)*	Autotétraploïde	Chaleur	Zhang *et al.* , 2010
Orge *(Hordeum vulgare)*	Autotétraploïde	Sécheresse	Chen and Tang, 1945
Tabac *(Nicotiana benthamiana)*	Octoploïde	Sécheresse, froid, déficit en nutriments	Deng *et al.* , 2012
Oranger trifolié *(Poncirus trifoliata)*	Autotétraploïde (porte-greffe)	Stress salin combiné à la sécheresse	Saleh *et al.* , 2008
Patate douce(*Ipomoea batatas*)	Hexaploïde	Sécheresse	Arisha *et al.* , 2020
Goji noir *(Lycium ruthenicum)*	Autotétraploïde	Sécheresse	Rao et *al.* , 2020
Trèfle *(Trifolium ambiguum)*	Hexaploïde	Sécheresse	Williams *et al.* , 2019
Blé *(Triticum species)*	Allotétraploïde/allohexaploïde	Stress salin	Yang *et al.* , 2014
Melon *(Citrullus lanatus)*	Autotétraploïde	Stress salin	Zhu *et al.* , 2018
Riz *(Oryza sativa)*	Tétraploïde	Stress salin	Tu *et al.* , 2014
Peuplier *(Populus cathayana)*	Allotétraploïde	Stress salin	Qiu *et al.* , 2021

Tableau 4 : Liste de publications référençant une tolérance aux stress abiotiques conférée par l'état polyploïde chez différentes espèces de plantes.

En considérant l'adaptation au stress hydrique à titre d'exemple, il s'avère que les polyploïdes indépendamment de leur ascendance phylogénétique sont capables de conquérir des habitats plus secs que ceux occupés par des diploïdes de mêmes clades (Te Beest *et al.*, 2012; Gunn *et al.*, 2020). En cherchant les facteurs permettant cette adaptation, il a été mis en évidence que la polyploïdie entraîne des modifications de la transpiration, de l'efficacité de l'utilisation de l'eau, du taux de photosynthèse, de la phénologie et de la morphologie (Soltis and Soltis, 2014; De Baerdemaeker *et al.*, 2018). Cette diversité témoigne de l'importance du phénomène de polyploïdisation et de sa propension à impacter tous les niveaux de la physiologie de la plante.

Stress biotiques

Les plantes subissent constamment des stress biotiques divers, plus ou moins intenses : attaques d'agents pathogènes, consommation par les espèces herbivores, compétition avec d'autres espèces. Dans ce contexte de pression constante, les polyploïdes parviennent à mieux résister ou tolérer les dommages que les espèces diploïdes. Un modèle propose que la polyploïdie augmente la résistance dans les systèmes gène-à-gène qui impliquent de nombreuses interactions hôte-pathogène (Oswald and Nuismer, 2007). C'est le cas pour des polyploïdes synthétiques issus d'un cultivar de pomme, *Malus domestica*, qui montrent une résistance accrue à l'agent fongique de la tavelure du pommier, *Venturia inaequalis*, par rapport aux cultivars diploïdes (Hias *et al.*, 2018). Les résistances concernent divers types de pathogènes ou prédateurs : des tétraploïdes synthétiques de la pomme de terre *Plectranthus esculentus* sont plus résistants aux nématodes à galles que les diploïdes (Hannweg *et al.*, 2016), tandis que des allotétraploïdes synthétiques de tabac présentent une capacité absente chez les ascendants diploïdes à résister à l'attaque de la larve du papillon sphinx (Anssour and Baldwin, 2010).

Face aux stress biotiques et abiotiques, l'effet de la polyploïdie est comparable au phénomène d'hétérosis qui correspond aux gains de vigueur d'une descendance hybride en comparaison des lignées parentales (Birchler *et al.*, 2010) et les études de l'un et l'autre phénomène peuvent être complémentaires pour expliquer les bases moléculaires des gains observés (Washburn and Birchler, 2014).

3.3.4 Succès évolutif et écologique des polyploïdes

La polyploïdie se manifeste à la fois comme un effet du stress environnemental, décrit précédemment, et comme une adaptation à celui-ci. Lorsqu'un ou plusieurs stress majeurs s'appliquent, il serait donc logique d'assister à l'émergence et à l'expansion d'espèces polyploïdes. Dans cette optique, l'épisode, il y a 66 millions d'années, de l'extinction massive des espèces lors de la transition du Crétacé au Paléogène qui a frappé les dinosaures mais également de nombreuses espèces végétales, est particulièrement marquant. La période, qui se caractérise par une longue phase d'obscurité et de températures basses (Renne *et al.*, 2015), correspond à une explosion du nombre d'espèces d'angiospermes polyploïdes (Vanneste, Maere, *et al.*, 2014; Lohaus and Van de Peer, 2016) qui prennent alors le pas sur les espèces gymnospermes (Condamine *et al.*, 2020) (Figure 20).

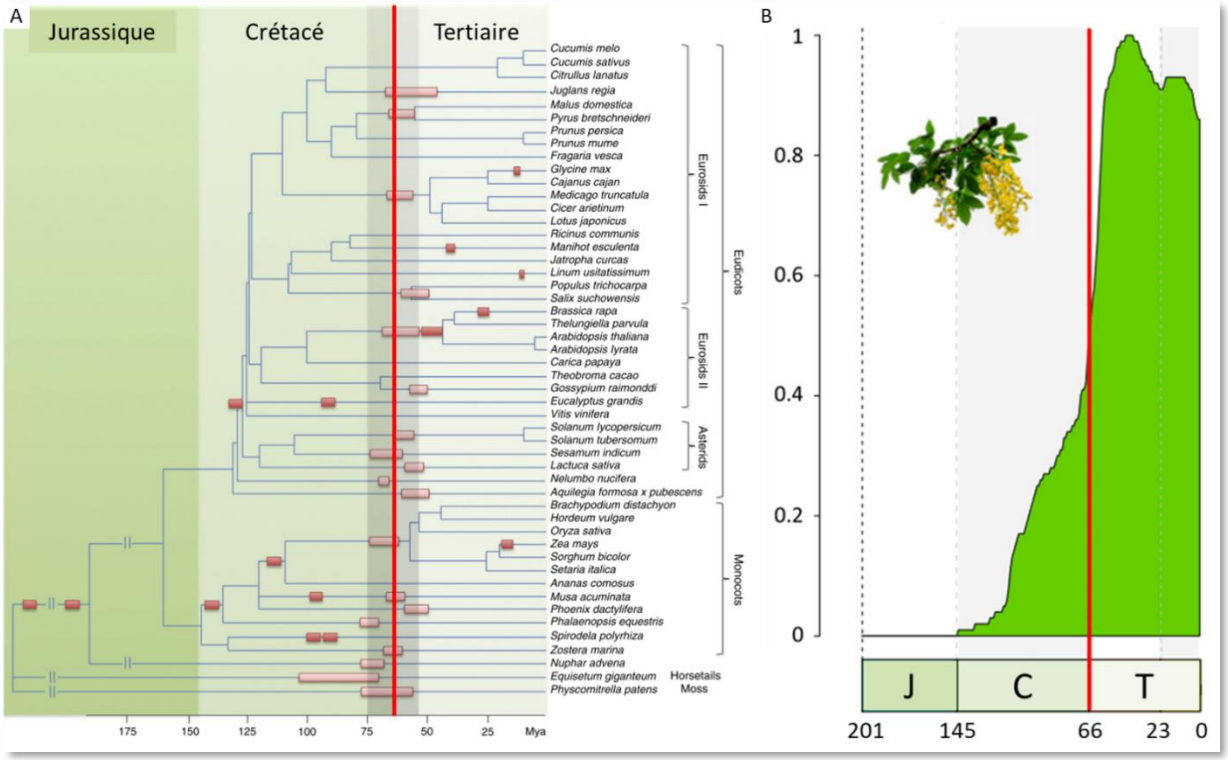


Figure 20 : Explosion des évènements de WGD lors de la transition entre Crétacé et Tertiaire il y a 66 millions d'années. A. Représentation des intervalles de temps associés à des évènements de WGD sur l'arbre évolutif des plantes terrestres depuis l'ère Jurassique (adapté de Vanneste et al., 2014). B. Evolution de la diversité des angiospermes sur la même période montrant l'essor des angiospermes au cours du Crétacé, illustré par le taux de fossiles d'angiospermes sur l'ensemble des fossiles des plantes vasculaires (Condamine et al., 2020).

Sur le plan de la répartition géographique contemporaine, la fréquence de la polyploïdie augmente de l'équateur vers les pôles, avec l'altitude et dans des zones arides comme le centre de l'Australie (Rice *et al.*, 2019). Cette répartition traduit le fait que les espèces polyploïdes sont plus performantes que leurs homologues diploïdes pour conquérir des habitats altérés ou vacants, à l'instar de zones urbanisées colonisées par un génotype tétraploïde d'orties (Rejlová *et al.*, 2019) ou de la zone arctique où les polyploïdes sont surreprésentés (Brochmann *et al.*, 2004). Ces études mettent en évidence des facteurs prédéterminant la réussite des polyploïdes, tels que la présence de stress biotiques notamment la sécheresse et la salinité des milieux, et d'autres de nature à favoriser leur établissement, tels qu'une concurrence réduite au sein d'une niche écologique ou le fait qu'ils aient un cycle pérenne leur procurant davantage de temps pour trouver des partenaires compatibles et produire une descendance fertile.

Outre les avantages liés à l'effet « gigantisme » et à la faculté à résister aux stress, plusieurs études mettent en lumière la plasticité phénotypique augmentée des polyploïdes relativement aux diploïdes (Hahn *et al.*, 2012; Yu *et al.*, 2021; Rejlová *et al.*, 2019; Wei *et al.*, 2019; Mattingly and Hovick, 2021) comme facteur déterminant leur adaptation dans des environnements instables, comme en témoigne leur surreprésentation au sein des plantes invasives. Cette surreprésentation est étudiée par Pandit *et al., e*n comparant les nombres de chromosomes et les niveaux de ploïdie d'espèces végétales menacées et d'espèces invasives. Ils montrent que pour l'ensemble des plantes considérées, l'augmentation du nombre de chromosomes est associée à une augmentation significative de la probabilité d'être envahissant et que les hauts niveaux de ploïdie sont associés avec une diminution significative de la probabilité d'être en danger. La même étude en prenant en compte la parenté des espèces du panel, montre que le fait d'être polyploïde plutôt que diploïde entraîne une diminution significative de la probabilité d'être en danger et une augmentation significative de la probabilité d'être une espèce invasive (Pandit *et al.*, 2011). La fréquence de la polyploïdie a également été analysée en rassemblant les données disponibles sur les niveaux de ploïdie de 128 espèces végétales invasives, présentes et considérées comme telles dans, au moins, trois régions du monde différentes. Parmi celles-ci, 70% des espèces considérées sont polyploïdes, ce qui excède les proportions de polyploïdes globalement observées quel que soit le biotope considéré (Rice *et al.*, 2019), et 43% d'entre elles présentent plusieurs niveaux de

ploïdie (Te Beest *et al.*, 2012). A contrario, la polyploïdisation a pu être considérée comme un 'cul de sac' évolutif dans les environnements les plus stables (Comai, 2005; Mayrose *et al.*, 2011; Arrigo and Barker, 2012), quoique ce point de vue soit encore discuté (Soltis *et al.*, 2014).

3.4 Impacts de la polyploïdie sur la structure et la régulation du génome

L'évènement de polyploïdie déclenche un ensemble de mécanismes qui se traduisent par des changements de la structure et de la régulation du génome au cours des générations successives post-duplication. Ces mécanismes sont interconnectés mais seront présentés de la manière suivante dans la prochaine section du manuscrit : impact de la polyploïdie sur la structure des génomes, incluant les modifications caryotypiques et les réarrangements géniques ; impact de la polyploïdie sur la régulation des génomes, incluant les effets sur la régulation épigénétique (méthylomique) et sur l'expression des gènes (transcriptomique).

3.4.1 Impacts de la polyploïdie sur la structure du génome

3.4.1.1 Modifications caryotypiques post-polyploïdie

Les modifications caryotypiques sont présentées ici en trois catégories. Les modifications locales qui interviennent au sein d'un unique chromosome constituent la catégorie des variations structurales. La catégorie des réarrangements chromosomiques désigne les évènements impliquant des échanges entre chromosomes. Enfin, les effets sur la dynamique des éléments transposables induits par la polyploïdie constituent la dernière catégorie d'évènements présentés.

Variations structurales

Les variations structurales peuvent correspondre à des évènements (i) de gain et perte de segments d'ADN (*Presence/Absence Variation* PAV), (ii) de duplications (*Copy Number Variation* CNV), et (iii) d'inversions (Figure 21.A). Les 3 mécanismes principaux sont :

- Le crossing-over inégal : lors de la méiose, les chromosomes homologues s'apparient sur les régions semblables, et peuvent éventuellement s'échanger. Ainsi, un segment d'ADN de l'un des deux chromosomes homologues est transféré à l'autre. Si cette portion de chromosome porte un

ou plusieurs gènes, il en résulte une perte de gènes (PAV) sur un chromosome et une duplication de gènes (CNV) en tandem sur l'autre.

- L'échange ectopique : lors d'une cassure double brin, il peut y avoir recombinaison avec un site non homologue, puis élongation de la molécule d'ADN suivant la matrice d'ADN recombinant, recopiant ainsi une portion du génome après la cassure. La réparation de la cassure a ensuite pour conséquence l'intégration de cette région dupliquée au génome induisant un CNV.

- Un gain médié par un élément transposable : lors de la rétrotransposition d'un élément transposable, il peut y avoir transcription d'une partie de la région chromosomique à proximité de l'élément transposable, comprenant éventuellement un ou plusieurs gènes. Lorsque l'ARN est rétrotranscrit au sein du génome, le ou les gènes transcrits par accident sont alors copiés dans une autre région du génome.

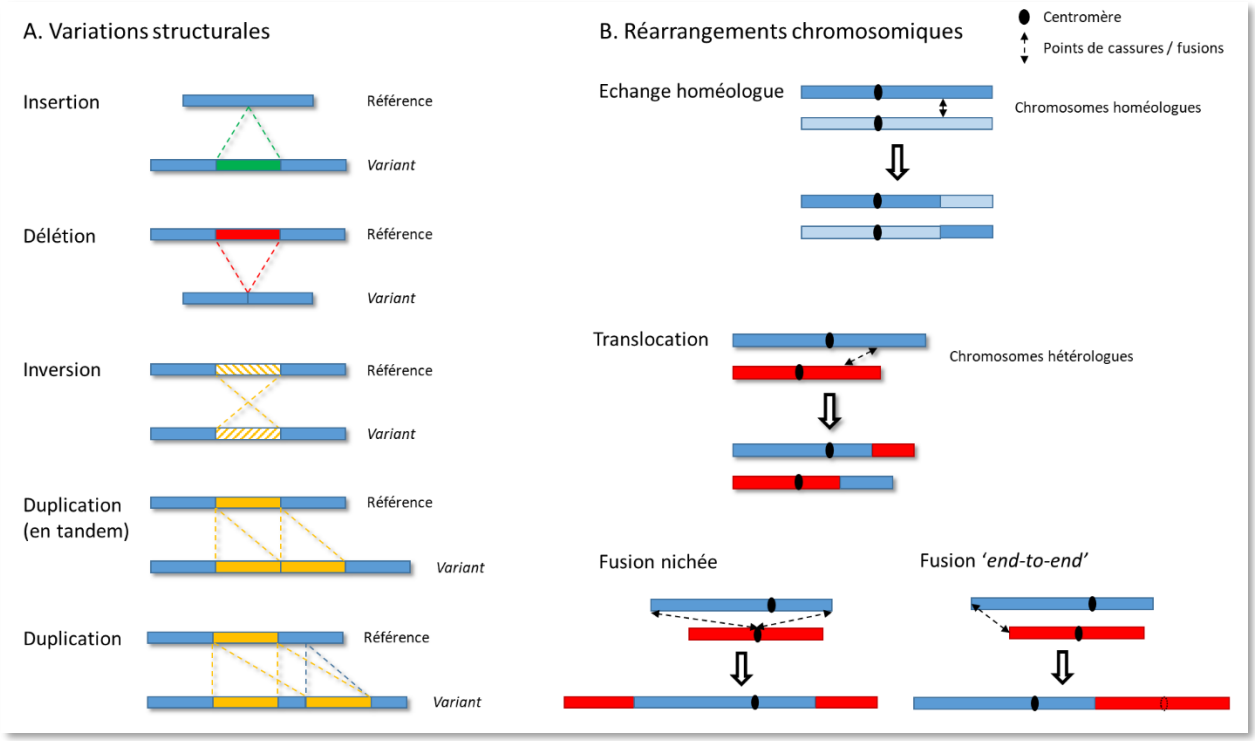


Figure 21 : Représentation canonique (A) des Variations structurales observables entre un génome de référence et un génome variant et (B) des réarrangements chromosomiques impactant un caryotype à l'échelle de chromosomes entiers (disploïdie).

Au-delà des PAV et des CNV, les inversions qui constituent une autre source de variation structurale, sont dues à deux cassures sur le même chromosome, suivies par un double

recollement après inversion du segment intermédiaire. Elles sont dites péricentriques si le centromère est compris dans le segment intermédiaire, ou paracentriques si les deux cassures se sont produites sur le même bras chromosomique.

L'ensemble de ces mécanismes lorsqu'ils concernent des gènes vont affecter leur régulation et leur expression, et sont donc susceptibles de modifier *in fine* les caractères exprimés par les plantes comme le montrent les multiples exemples relatés dans la littérature (Gabur *et al.*, 2019; Zhou and Gaut, 2020). Comme évoqué précédemment (page 29), la possibilité de disposer de plusieurs génomes de haute qualité à comparer les uns aux autres, entraîne le développement de nouveaux logiciels dédiés à la détection et à la caractérisation des variations structurales (Yuan *et al.*, 2021).

Réarrangements chromosomiques

Les réarrangements chromosomiques (Figure 21.B) correspondent à des échanges homéologues, ou des fusions et fissions des chromosomes susceptibles de modifier la structure des caryotypes en modifiant le nombre et la taille des chromosomes.

- Les échanges homéologues sont des échanges de fragments chromosomiques entre deux chromosomes homéologues.

- Les fusions et fissions des chromosomes résultent de cassures et de réparations illégitimes. Les fusions peuvent se faire au niveau des télomères, 'end-to-end', ou entre un centromère et des télomères, fusion nichée ou *nested chromosomes fusion*.

Ces réarrangements chromosomiques post-polyploïdie, en modifiant le nombre et la taille des chromosomes, constituent un mécanisme majeur permettant aux néopolyploïdes de différencier rapidement les chromosomes homéologues des chromosomes homologues pour assurer des méioses équilibrées (Schubert and Vu, 2016; Mandáková and Lysak, 2018). L'accumulation de réarrangements chromosomiques et de variations structurales tend ainsi au cours de l'évolution à rétablir une structure caryotypique et un fonctionnement totalement diploïde.

Effets sur la dynamique des éléments transposables

Les éléments transposables (ET) sont des éléments génétiques mobiles qui représentent une fraction, variable mais fréquemment importante, des génomes végétaux. Leur capacité à proliférer en se multipliant dans les génomes et, à l'opposé, l'existence de mécanismes conduisant de façon très efficace à leur élimination en font des moteurs majeurs de l'expansion et de la contraction des génomes au fil de l'évolution des plantes terrestres (Hidalgo *et al.*, 2017). Au-delà des variations qu'ils induisent sur la taille des génomes, la présence et la mobilité des ET ont de multiples impacts. D'une part, les transpositions induisent de nouvelles insertions, qui dans la plupart des cas seront sélectivement neutres ou délétères, mais dans certains cas pourront produire un avantage sélectif (Arkhipova, 2018). D'autre part, la nature répétée des ET offre de nombreuses possibilités d'appariements de séquences susceptibles de provoquer des recombinaisons, et en cela, ils constituent une force créatrice des variants structuraux présentés précédemment qui peuvent concerner des gènes fonctionnels situés à proximité physique des ET. Enfin, les ET sont des acteurs cruciaux des mécanismes de régulation des gènes en étant, à la fois, une source importante de promoteurs et d'éléments de régulation transcriptionnelle, mais également, la cible de modifications épigénétiques contrôlant leur activité, aussi appelées *silencing*, dont la méthylation susceptible d'affecter également les gènes à proximité et de modifier leur régulation (Lisch, 2009).

Dans le contexte de la polyploïdisation, la fusion de deux génomes revient à combiner deux populations différentes d'ET et de siRNA qui les ciblent dans un seul génome. Ce choc peut déclencher l'activation des ET, entraîner des modifications épigénétiques des ET eux-mêmes et des gènes voisins et, par conséquent impacter la régulation des uns et des autres (Vicient and Casacuberta, 2017). Cependant, il ne semble pas y avoir un mécanisme unique et partagé. Si l'activation de la transcription des ET a été observée chez des polyploïdes synthétiques d'*Arabidopsis* (Madlung *et al.*, 2005), des blés polyploïdes (Kashkush *et al.*, 2003) et le café allopolyploïde (Lopes *et al.*, 2013), *a contrario,* aucune modification significative de la teneur en ET n'a été montrée après la polyploïdisation de *Arabidopsis arenosa* (Baduel *et al.*, 2019) et il a été montré une augmentation des siRNA liés aux ET chez des spartines allopolyploïdes suggérant un renforcement de la répression de leur transcription en réponse à la polyploïdisation (Cavé-

Radet *et al.*, 2019). Mais même en l'absence d'augmentation de leur transcription, la polyploïdie est susceptible d'entraîner une augmentation du contenu en ET du fait du relâchement de la sélection purificatrice au niveau des loci dupliqués (Baduel *et al.*, 2019). En dépit de ces résultats a priori contradictoires, l'hypothèse que les différences de contenus en ET entre les deux progéniteurs d'un organisme allopolyploïde soient à l'origine de l'expression différentielle des gènes homéologues est prévalente. Les gènes dont l'expression est inhibée du fait de la proximité d'ET insérés post-polyploïdie pourraient subir une évolution accélérée de leurs séquences aboutissant à leur élimination (Bottani *et al.*, 2018; Alger and Edger, 2020). La question de l'évolution des paires de gènes homéologues est traitée dans la section suivante.

3.4.1.2 Modifications géniques post-polyploïdie

Définitions et nomenclature

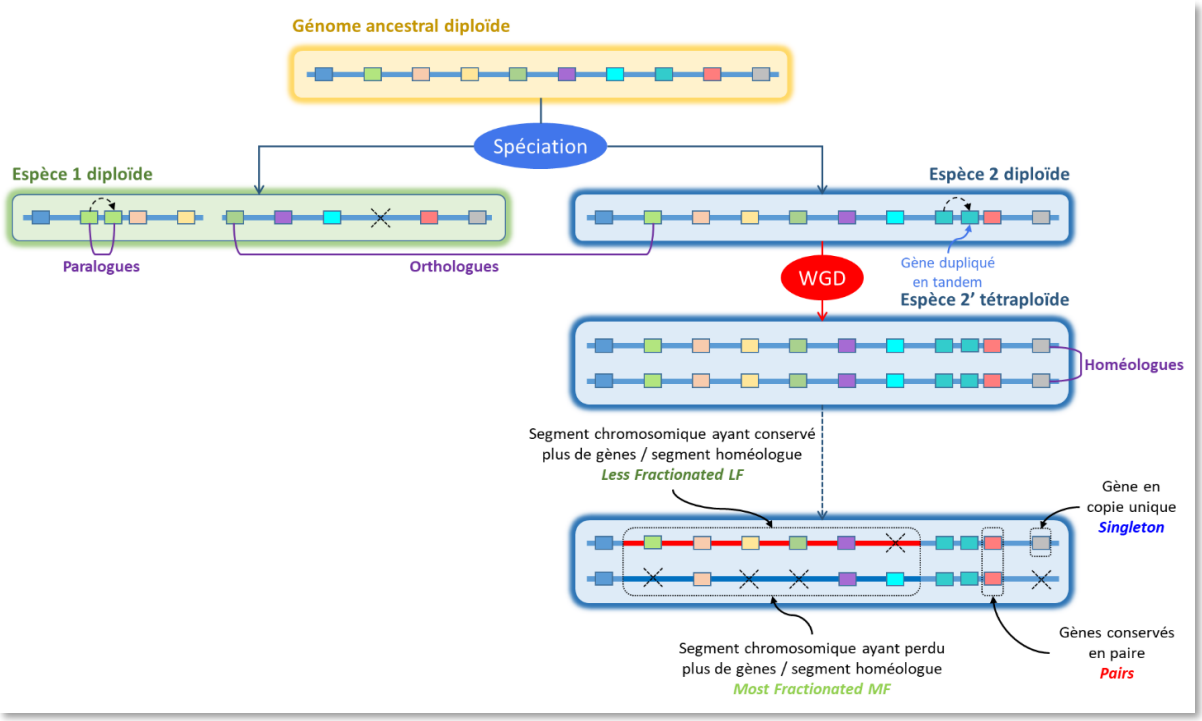


Figure 22 : Nomenclature des relations entre gènes (boîtes colorées) et définition des blocs post WGD. Paralogues : gènes issus d'une duplication. Orthologues : gènes descendants d'un gène ancestral par spéciation. Homéologues : gènes issus d'un gène ancestral par WGD.

Les gènes issus d'un évènement de duplication, quel qu'il soit, sont des gènes paralogues. Les gènes provenant d'un gène ancestral commun sont des gènes homologues. Les gènes homologues entre deux espèces sont des gènes orthologues. Les gènes homologues au sein d'une espèce allopolyploïde conservant la structure des progéniteurs sont des gènes homéologues. Les gènes homologues entre espèces issues d'un évènement de polyploïdisation précédent la spéciation et conservés en paires sont dits ohnologues. (Glover *et al.*, 2016; Fitch, 1970). A la suite d'un événement de polyploïdie, chaque gène ancestral va produire une paire de gènes. Ces deux gènes auront chacun un destin propre. Ils pourront être conservés en paire ou, si l'une des deux copies est éliminée, redevenir un gène en copie unique, ou *singleton*.

A l'échelle des chromosomes homéologues, les pertes de gènes peuvent être équilibrées entre les sous-génomes post-polyploïdie, on parle alors de fractionnement équilibré. A l'inverse, les pertes de gènes peuvent être inégales le long des chromosomes homéologues, on parle alors de fractionnement asymétrique, ou *biased fractionation.* Dans ce cas, le segment chromosomique ayant conservé plus de gènes sera dit *Less Fractionated* (LF), tandis que le segment ayant perdu plus de gènes sera dit *Most Fractionated* (MF) (Figure 22).

Si le fractionnement opère préférentiellement sur un des sous-génomes issus de la WGD, les fractions LF et MF correspondent alors aux chromosomes entiers des sous-génomes, c'est la dominance des sous-génomes. Au sein des génomes modernes, cette dominance se définit par une différence de nombre de gènes orthologues (à partir d'une espèce proche non dupliquée) ou ancestraux (à partir de la reconstruction d'un génome ancestral pre-duplication) portés par les blocs chromosomiques dupliqués (ou sous-génomes) : on parle de dominance structurale des sous-génomes post-polyploïdie (Figure 23).

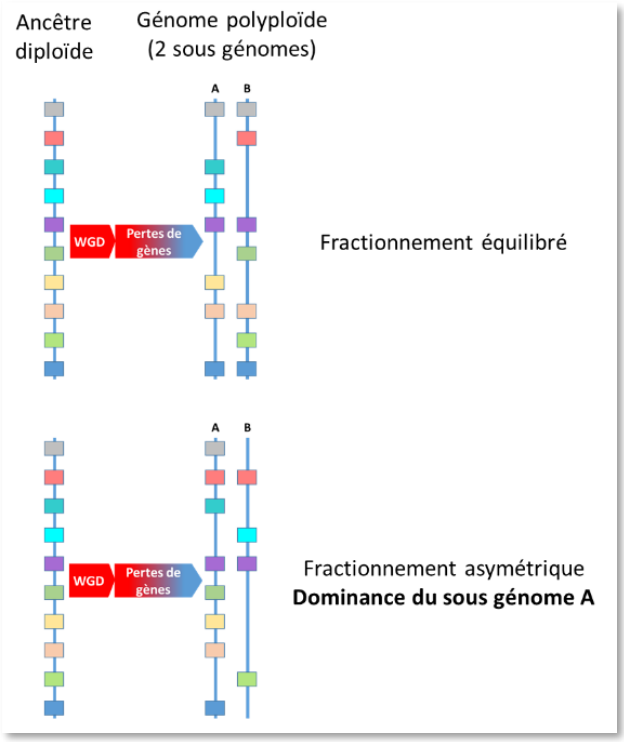


Figure 23 : Représentation des deux types de fractionnement. Suite à la WGD, la perte des gènes peut se faire de façon équilibrée (haut) ou de façon asymétrique (bas) Dans ce cas, le sous-génome A conserve plus de gènes, il est dominant et le sous-génome B perd le plus de gènes et est sensible.

Ce phénomène de dominance des sous-génomes a initialement été observé chez *Arabidopsis thaliana* (Thomas *et al.*, 2006) puis chez le maïs (Woodhouse *et al.*, 2010), avant d'être étudié et mis en évidence chez de nombreuses espèces (Tableau 5).

Taxon / taxa	Date de la WGD (millions d'années)	*'biased fractionation'*	Dominance	Références
Zea mays	8	oui	oui	(Schnable et al., 2011)
Glycine max	13	non	non	(Garsmeur et al., 2014)
Cucurbita sp.	3–26	non	non	(Sun et al., 2017)
Brassica rapa	15	oui	oui	(Mandáková et al., 2017)
Arabidopsis thaliana	47	oui	oui	(Thomas et al., 2006)
Medicago sativa	58	oui	N/A	(Garsmeur et al., 2014)
Gossypium sp.	60	oui	oui	Renny-Byfield et al., 2015)
Musa acuminata	65	non	non	(D'hont et al., 2012)
Populus trichocarpa	65	non	non	(Garsmeur et al., 2014)
Poaceae	70	oui	oui	(Garsmeur et al., 2014)

Tableau 5 : revue des génomes de plantes polyploïdes pour lesquels le fractionnement et les relations de dominance des sous génomes ont été étudiés. D'après Wendel et al., 2018. Références : Schnable et al., 2011; Garsmeur et al., 2014; Sun et al., 2017; Mandáková et al., 2017; Thomas et al., 2006; Renny-Byfield et al., 2015; D'hont et al., 2012

Lorsqu'il n'y a pas de biais de pertes de gènes entre sous-génomes mais des différences en termes d'expression, l'un des sous génomes contribuant majoritairement au transcriptome, on parle de dominance d'expression des sous-génomes post-polyploïdie (Rapp *et al.*, 2009).

3.4.2 Impacts de la polyploïdie sur la régulation du génome

3.4.2.1 Expression

Les modifications du niveau d'expression des gènes sont une conséquence immédiate de la polyploïdie au sein de la cellule. Elles peuvent suivre plusieurs patrons (Figure 24). Le niveau d'expression chez le polyploïde peut correspondre à la moyenne des niveaux d'expression des progéniteurs. Dans ce cas, les deux copies présentes s'expriment à leur niveau originel au sein d'un transcriptome dont la 'taille' est multipliée par deux sous l'effet de la WGD, on parle alors d'additivité. Le niveau d'expression du polyploïde peut également être différent de la moyenne des niveaux d'expression des progéniteurs. Il peut soit être aligné sur celui d'un des deux gènes homéologues, on parle alors de dominance d'un sous-génome, soit ne pas être corrélé avec les niveaux d'expression des progéniteurs, on parle alors de transgressivité soit en sous-expression ou en surexpression. L'étude des patrons d'expression pour un panel de quatre espèces polyploïdes, coton, café, *Tragopogon* et blé, indique qu'une proportion de 19 à 61% des gènes présente des patrons d'expression non-additifs (Yoo *et al.*, 2014).

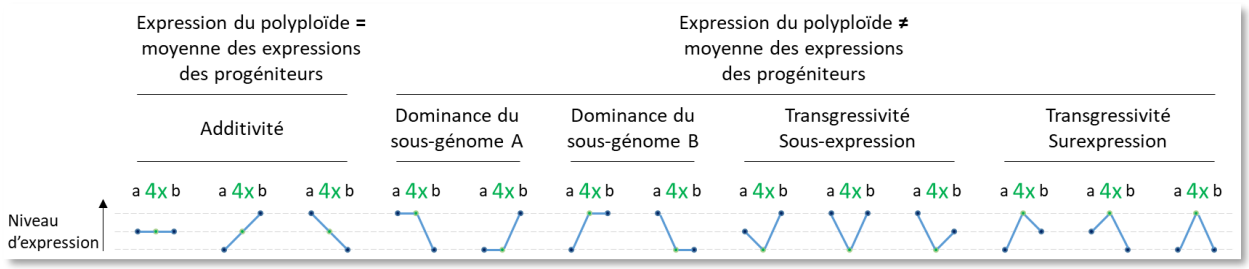


Figure 24 : Patrons d'expression des gènes chez les progéniteurs diploïdes (notés a et b) et leur descendant allotétraploïde (noté 4x).

Alors que l'additivité procède d'un maintien des régulations des progéniteurs au sein du polyploïde, la dominance et la transgressivité résultent de modifications de celles-ci. Plusieurs facteurs peuvent impacter les mécanismes de régulation de l'expression. La différence de contenus en ET entre les progéniteurs (précédemment évoquée) est une cause reconnue de différenciation de l'expression entre les copies homéologues chez le polyploïde, parfois présenté

comme un facteur majeur (Bottani *et al.*, 2018) ou comme une cause potentielle parmi d'autres (Yoo *et al.*, 2014). Une autre cause expliquant l'expression non additive est la mise en place de nouvelles régulations. Il existe deux types de régulations de l'expression des gènes, soit en *cis* lorsqu'une séquence régule l'expression d'un gène présent à proximité sur le même chromosome, soit en *trans* lorsque l'expression d'un gène est régulée par un élément agissant à distance de la séquence qui le code. La polyploïdisation a pour effet de multiplier les cibles de chaque facteur régulant en *trans* et par conséquent de multiplier les possibilités de régulations. De nouvelles régulations peuvent se mettre en place en fonction d'effets doses, de changements de méthylation ou si les séquences de deux régions régulatrices divergent post-polyploïdisation (Figure 25).

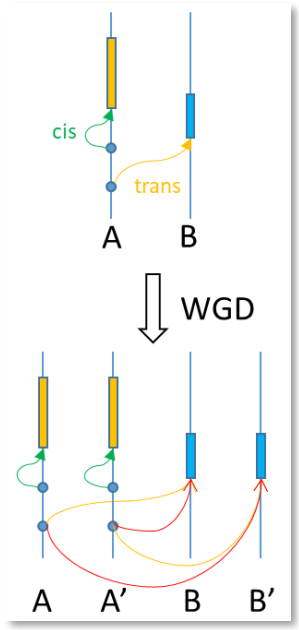


Figure 25 : Mise en place de nouvelles relations de régulation en trans suite à une WGD. La duplication des séquences régulatrices doubles le nombre d'interaction en trans. Les nouvelles voies de régulation post-polyploïdie sont illustrées par les flèches rouges.

Au-delà du modèle représenté, la présence de deux sets de chromosomes homéologues modifie les interactions de la chromatine dans le noyau ce qui multiplie les nouvelles interactions potentielles entre des régions chromosomiques qui n'étaient jusqu'alors pas en contact. Plusieurs études suggèrent que cette complexification des voies de régulations chez les polyploïdes relativement aux diploïdes contribue à générer les nouvelles régulations précédemment présentées (Bao *et al.*, 2019; Hu and Wendel, 2019). Une explication également évoquée est

l'influence du contenu cytoplasmique maternel : le fait que chez les angiospermes les génomes des organelles soient transmis par le progéniteur maternel serait à l'origine d'une plus grande compatibilité avec le génome nucléaire maternel conférant une dominance de l'expression des homéologues issus de ce dernier (Yoo *et al.*, 2013).

3.4.2.2 Epigénétique

Les effets de la polyploïdisation sur le paysage épigénétique demeurent moins étudiés que ses impacts sur la perte de gènes et sur leur expression (Soltis *et al.*, 2016) en dépit de l'importance de la méthylation sur le fonctionnement des génomes à de multiples niveaux (Figure 26).

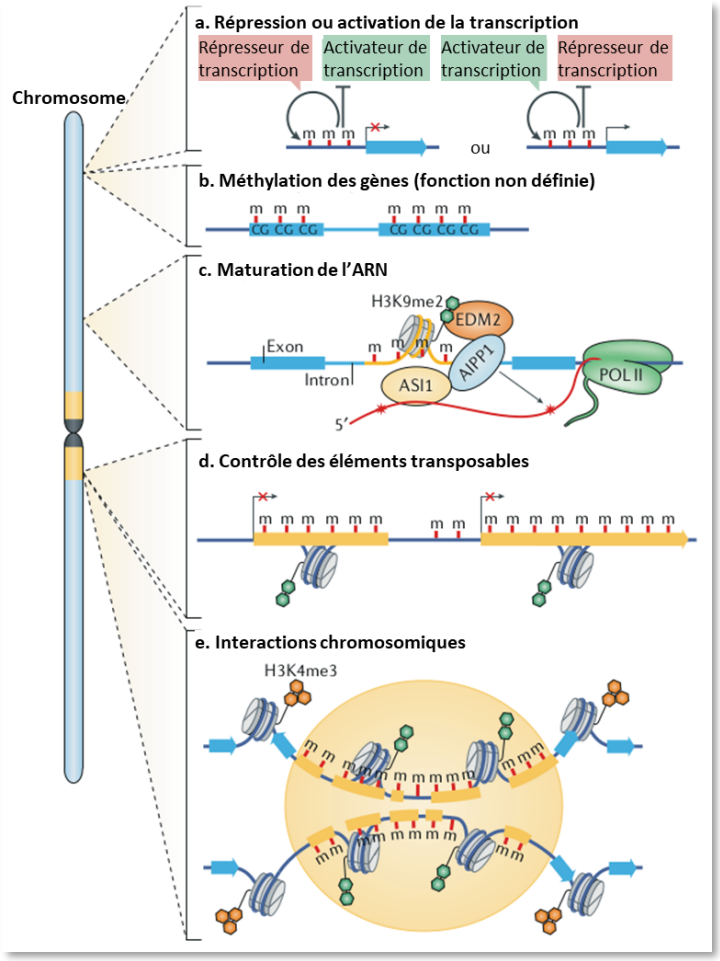


Figure 26 : Différents niveaux d'impacts de la méthylation d'ADN. L'ADN peut être méthylé au niveau des promoteurs (a), des gènes (b), des introns (c), des éléments transposables (d) ou au niveau des séquences intergéniques. Ces différentes localisations impactent la régulation des gènes (a) et (b), la maturation de l'ARN (c), le contrôle de l'activité des ET (d) et les interactions chromosomiques (e) selon des modalités plus ou moins connues. D'après Zhang et al., 2018.

La méthylation des cytosines, qui constitue une des marques épigénétique majeure, est notamment reconnue comme un facteur déterminant de la régulation de l'expression des gènes et du contrôle des éléments transposables, susceptible de se mettre en place à une plus haute fréquence que la mutation de la séquence nucléotidique. L'étude des ET évoquée précédemment a montré que la WGD induit des changements de méthylation des gènes à proximité des ET méthylés (Woodhouse *et al.*, 2014), mais l'influence directe de la méthylation sur l'expression des gènes étant quantitative et non-linéaire, il est difficile jusqu'à présent de tirer des conclusions claires (Diez *et al.*, 2014; Zhang *et al.*, 2018).

Cependant des études tentent de détisser les liens complexes entre méthylation, ET et polyploïdie. L'observation de l'évolution de la méthylation d'une espèce de riz tétraploïde synthétique montre, qu'après la WGD, la méthylation reste globalement équivalente au niveau de celle des progéniteurs diploïdes mais les niveaux de méthylation des sous-génomes tendent à se différencier (Li *et al.*, 2019). En ce qui concerne la méthylation des gènes issus d'une WGD, qu'ils soient conservés en paires ou qu'une unique copie soit conservée, les observations sont contradictoires. Chez le soja, tétraploïde relativement récent, les gènes conservés en paires sont plus méthylés et, en moyenne, plus fortement exprimés que les singletons. Les gènes méthylés accumulent également moins de mutations que les gènes non méthylés, ce qui suggère que les gènes conservés en paires évoluent plus lentement que les gènes conservés en copie unique (Kim *et al.*, 2015). A l'opposé, une étude sur le lotus montre les gènes qui perdent une copie homéologue et se retrouve en copies uniques présentent des niveaux d'expression et de méthylation élevés (Shi *et al.*, 2020). Ces résultats mettent en évidence la nécessité d'explorer davantage le sujet de la méthylation et de son influence sur l'expression chez les polyploïdes.

3.4.2.3 Impacts des changements structuraux et régulationnels sur les paires de gènes

Les impacts de la polyploïde, présentés ci-dessus, sont susceptibles de s'appliquer différentiellement au sein des paires de gènes. Une des copies pourra accumuler davantage de mutations, avoir une expression diminuée relativement à son niveau initial, présenter des changements de niveau de méthylation. L'apparition de tels changements ou, au contraire, le maintien des niveaux initiaux induiront divers chemins évolutifs pour les paires de gènes issues

de la duplication, de la conservation de deux copies fonctionnelles à la disparition complète de l'une des copies. Quatre scénarios sont observés.

- Redondance fonctionnelle

Il est possible que les deux copies d'un gène soient toutes deux maintenues dans le génome et continuent à exprimer la fonction ancestrale, conduisant à une redondance fonctionnelle (Vavouri *et al.*, 2008; Gout and Lynch, 2015). Plusieurs hypothèses, non-exclusives, existent pour expliquer les cas de conservation et de co-expression des deux copies homéologues. Un premier modèle explique cette possibilité en proposant que la réduction de l'expression puisse favoriser la rétention des deux copies conservant leur fonction ancestrale (Qian *et al.*, 2010). A l'opposé, l'hypothèse du dosage absolu, considérant que plus le gène est exprimé plus l'organisme bénéficie d'un avantage, explique la conservation de la paire par l'avantage direct conféré par la « double expression ». L'hypothèse du dosage relatif repose sur le fait que l'équilibre stœchiométrique soit important pour le maintien de la fonctionnalité de gènes impliqués dans certains réseaux ou complexes protéiques ce qui conduit à la conservation des deux homéologues (Yoo *et al.*, 2014). Chez la levure *S. cerevisiae*, il a été démontré que des gènes dupliqués maintenaient une redondance fonctionnelle pendant plusieurs millions d'années (Dean *et al.*, 2008).

- Sous-fonctionnalisation

Au sein d'une paire de gènes dupliqués, les deux copies sont conservées mais présentent des profils d'expression divergents, souvent complémentaires, c'est la sous-fonctionnalisation. La complémentarité peut provenir d'un changement dans les séquences régulatrices, conduisant les deux copies à avoir des patrons d'expressions spatiaux ou temporels différents mais complémentaires (Force *et al.*, 1999; Panchy *et al.*, 2016). C'est par exemple le cas chez la tomate pour deux gènes, PhyB1 et PhyB2 constituant une paire de gènes issue d'une WGD, appartenant à la famille des phytochromes, des récepteurs sensibles à la lumière jouant un rôle dans le développement des plantes. PhyB1 et PhyB2 présentent des fonctions communes de régulation de la photosynthèse mais PhyB1 régule spécifiquement des réponses à la gravité et à l'auxine, ce

qui suggère qu'ils se sont progressivement sous-fonctionnalisés depuis leur duplication (Carlson *et al.*, 2020).

- Pseudogénisation

Après la duplication, les gènes de la paire nouvellement créée peuvent connaître différentes pressions de sélection. Si une seule une copie du gène continue à être soumise à des contraintes sélectives purificatrices du fait de l'utilité de l'expression sa fonction initiale, il est probable que l'autre copie n'étant plus sous sélection accumule des mutations délétères, jusqu'à être perdre complètement sa capacité à être exprimé et sa qualité de gène. Cependant, lors de l'accumulation progressive des mutations, le gène qui se mue en pseudogène peut être conservé dans le génome. C'est ainsi qu'*Arabidopsis* et le riz contiennent des milliers de pseudogènes dans leurs génomes (Zou *et al.*, 2009). Le pseudogénisation peut constituer une impasse évolutive et conduire à la disparation de la copie pseudogénisée mais peut également permettre l'émergence d'une nouvelle fonction, ou néo-fonctionnalisation (voir ci-dessous).

- Néo-fonctionnalisation

Dans une minorité des cas, les mutations accumulées par un pseudogène peuvent conférer une nouvelle fonction. Si cette fonction est avantageuse, elle sera soumise à des contraintes sélectives conduisant à sa fixation dans la population, ce qui constitue la néo-fonctionnalisation. Il existe deux modèles pour expliquer ce mécanisme. Le modèle de Dyhkhuzen-Hartl propose que les mutations au niveau du gène dupliqué soient fixées par dérive et que plus tard, lors d'un changement d'environnement, le nouveau gène devienne avantageux pour l'organisme (Dykhuizen and Hartl, 1981), alors que le "modèle d'adaptation" propose qu'une mutation adaptative se trouve immédiatement fixée car immédiatement avantageuse (Innan and Kondrashov, 2010). Une analyse transcriptomique du maïs a permis de déterminer que 13% des paires de gènes générées par la duplication la plus récente du génome ont été soumises à une néo-fonctionnalisation s'exprimant dans les feuilles (Hughes *et al.*, 2014).

3.5 Polyploïdie et néodiploïdisation

L'ensemble des mécanismes à l'œuvre post-WGD, incluant les réarrangements structuraux, la modification du paysage des ET et de la méthylation, les changements de l'expression, de la fonction de gènes voire leur perte, tendent finalement à rendre au génome un caractère diploïde. Au terme de ce processus appelé néodiploïdisation, un chromosome aura un unique homologue et n'aura donc plus d'homéologue identifiable sur la base de sa structure globale, la majorité des paires de gènes produites par la WGD aura subi un processus de néo-fonctionnalisation, de pseudogénisation ou de perte d'une des copies homéologues, et la taille de génome se trouvera réduite par rapport à la taille initiale du néopolyploïde. *Arabidopsis thaliana* illustre l'efficacité de ce processus en présentant un nombre réduit de chromosomes (5) et un génome de petite taille (120Mb) contenant relativement peu de gènes (environ 26 000). Ce génome compact a pourtant subi de multiples WGD. *Arabidopsis thaliana* est issu de l'ancêtre commun à partir duquel l'ensemble des dicotylédones se sont différenciées, il y a environ 130 millions d'années. Son génome a subi une triplication et deux duplications. La reconstruction des caryotypes ancestraux montre qu'il a évolué à partir d'un set de 7 protochromosomes qui a donné suite à la triplication γ un premier intermédiaire à 21 protochromosomes. Ce nombre se réduit à 16 protochromosomes chez l'ancêtre commun des malvidées. Suivent 2 nouvelles duplications spécifiques des Brassicacées, β et α, avant une réduction à 8 puis 5 protochromosomes (Murat, Louis, *et al.*, 2015). Cette histoire évolutive témoigne à la fois de l'efficacité et du caractère récurrent des mécanismes de réduction du nombre de chromosomes, de perte de gènes et de maîtrise de la taille du génome.

Ce processus cyclique est parfois représenté comme une suite d'évènements ordonnés, la perte de gènes entraînant un fractionnement biaisé suivi d'une réduction de la taille du génome et finalement de réarrangements chromosomiques, le tout résumé par les termes *wash-rinse-repeat* (Wendel, 2015). Mais, compte tenu de la complexité des mécanismes à l'œuvre et des incertitudes sur leur fonctionnement qui sont toujours l'objet de débats, il semble sage de considérer la néodiploïdisation comme le résultat d'un ensemble de processus interconnectés (Figure 27), à l'image du point de vue développé par Soltis et ses collaborateurs (2016). Ces derniers posent par ailleurs la question des connaissances manquantes pour proposer un scénario

explicitant les règles régissant le processus évolutif depuis la polyploïdisation jusqu'à la néodiploïdisation, en englobant la diversité des végétaux.

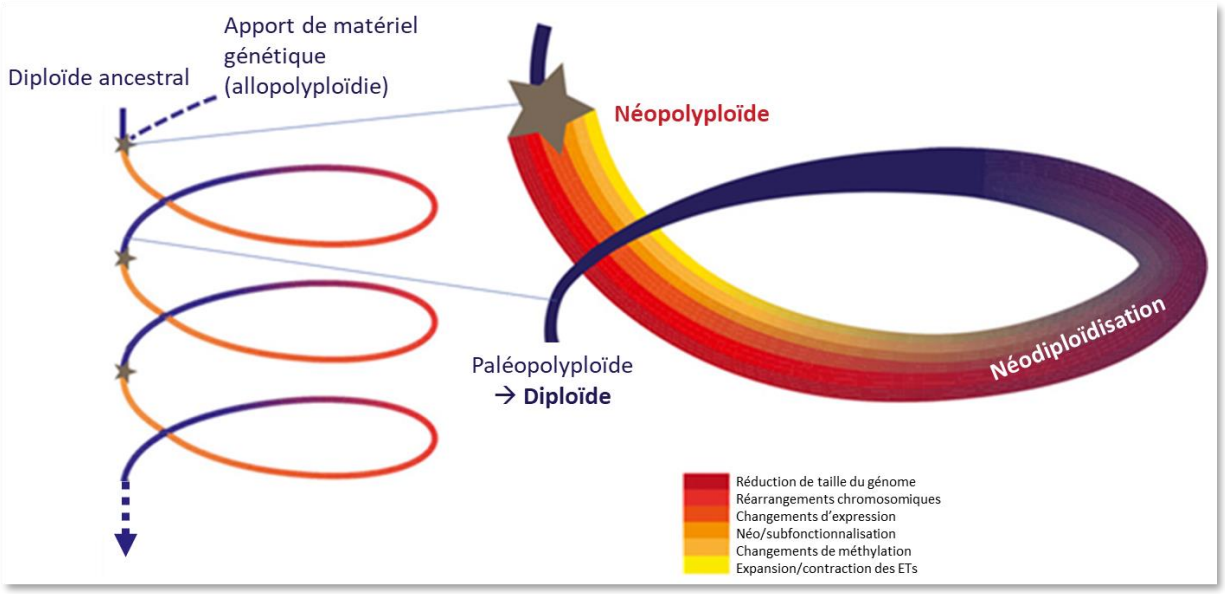


Figure 27 : Evolution cyclique du génome post-polyploïdie. Cette représentation du processus « wash, rinse, repeat » se veut nuancée en n'ordonnant pas les différentes forces évolutives entraînant le retour à l'état diploïde (Réduction de taille du génome, Réarrangements chromosomiques, Changements d'expression, Néo/subfonctionnalisation, Changements de méthylation, Expansion/contraction des ET), dans le but d'illustrer la complexité des mécanismes à l'œuvre et les incertitudes sur leur fonctionnement. Adapté de Soltis et al., 2016.

4 Questionnement scientifique de la thèse

L'importance des effets de la polyploïdie sur la capacité d'adaptation des plantes pose la question des mécanismes à l'œuvre lors de la reprogrammation génomique post-polyploïdie. Les nombreux travaux sur le sujet ont mis en évidence l'importance des différences entre progéniteurs diploïdes chez les allopolyploïdes, notamment en termes de contenu en ET, la nécessité pour les néopolyploïdes de mettre en place des solutions pour réaliser des méioses équilibrées, l'expression coordonnée des gènes homéologues, la capacité des polyploïdes à générer de la nouveauté génétique, le retour à l'état diploïde. Des mécanismes moléculaires pour expliquer ces phénomènes ont été avancés tels que le remodelage de la méthylation, le relâchement de la pression de sélection autorisant l'augmentation des taux de mutation, ou la dominance des sous-génomes. Mais ces mécanismes semblent varier d'une espèce à l'autre et les questions de leurs importances relatives et de leur séquence temporelle demeurent ouvertes. De même, face à la variété des réponses à la WGD, se pose la question du caractère stochastique des phénomènes et de leur récurrence. Pour répondre à ces questions, il faudra disposer :

- D'une analyse comparative impliquant plusieurs espèces dans une même étude pour tester la généricité des effets post-polyploïdie sur la structure et la régulation des génomes ainsi que des possibles mécanismes associés,

- De données omiques permettent de caractériser les effets post-polyploïdie en termes d'expression, d'épigénétique, de mutations de la séquence et de la structure de chromosomes,

- De méthodes et d'outils communs pour intégrer et comparer ces données chez différentes espèces pour exclure les biais d'interprétation,

- De connaissances bibliographiques par espèces pour, dans un second temps, comparer les résultats obtenus pour chaque espèce.

L'objet de ce travail de thèse est de contribuer à la compréhension des modifications génomiques post-polyploïdie, par une étude comparative chez les *Poaceae*, famille présentant plusieurs évènements de polyploïdisation depuis 100 millions d'années, en se basant sur la reconstruction des génomes ancestraux et en intégrant un large panel de données omiques.

II. RESULTATS

1 Objectifs

Mon travail de thèse, dont les résultats sont présentés dans ce chapitre, a pour but de réaliser une étude intégrée des effets de la polyploïdie chez huit espèces de graminées à travers la comparaison et l'intégration de données omiques, en s'appuyant sur la reconstruction des génomes ancestraux. La reconstruction des génomes ancestraux post- et pré-duplication permet, d'une part, d'identifier les gènes ancestraux qui ont été dupliqués au cours de l'évolution, pour mettre en évidence les variations structurales et les pertes de gènes, et d'autre part, de comparer les dynamiques de mutations, d'expression et de méthylation des régions et des gènes dupliqués.

Les questions majeures adressées dans ce travail sont :

Aspect structural :

- Quels gènes sont retenus ou perdus après un événement de polyploïdie ? Quelles sont leurs fonctions biologiques ?

- La perte des gènes après un évènement de polyploïdie suit-elle le principe de dominance des sous génomes ? Quelles sont les fonctions biologiques des gènes portés par les blocs LF et MF ?

Aspect régulationnel :

- Quel est le devenir de la régulation des gènes et des régions dupliquées en termes de taux de mutation, de niveaux d'expression, de fréquences de méthylation et de dynamiques à l'échelle populationnelle (SNP) selon (i) qu'ils appartiennent à des régions LF ou MF, (ii) qu'ils soient conservés en paires ou en copie unique, et (iii) entre les deux copies de gènes dupliqués ?

- Quelles sont les relations entre méthylation et expression des gènes ?

Aspect modélisation :

- Est-il possible de donner un sens à l'analyse intégrée de l'ensemble de ces données par une approche multiomiques ?

- *In fine*, est-il possible, à partir de l'ensemble de ces résultats de proposer un scénario évolutif de la reprogrammation génomique post-polyploïdie pour les céréales identifiant les forces et évènements majeurs et leur séquence dans le temps ?

2 Concepts et éléments de contextes

2.1 Les céréales, modèle pour l'étude des processus évolutifs

2.1.1 Histoire et socioéconomie

Le terme céréale désigne des plantes monocotylédones de la famille de *Poaceae*. Cette famille comprend le maïs, le blé et le riz qui sont des espèces majeures pour l'alimentation humaine depuis la fin de la préhistoire. Elles ont toutes trois été domestiquées depuis 10 000 ans, indépendamment dans trois lieux distincts comptant parmi les 6 principaux centres de domestication des plantes répertoriés sur le globe (Gepts, 2004). Le maïs a été domestiqué au Mexique dans la vallée du Rio Balsas (Piperno *et al.*, 2009), les *Pooideae, blé et orge,* dans la zone du croissant fertile situé au proche Orient entre le Nil et le Tigre (Brown *et al.*, 2009; Pont *et al.*, 2019) et le riz au sud-est de la Chine dans la vallée du fleuve Yangzi Jiang (Kovach *et al.*, 2007). La domestication des céréales constitue un fait central, cause ou conséquence majeure selon les interprétations, de la révolution néolithique, qui voit l'être humain cueilleur chasseur se sédentariser pour devenir cultivateur, il y a entre 10 000 et 5 000 ans (Mazoyer and Roudart, 1997).

Aujourd'hui, les principales céréales en termes de production sont le maïs, le blé, le riz, les mils (y compris le sorgho), l'orge et l'avoine (par ordre de tonnage décroissant). Elles sont cultivées pour leurs grains et utilisées pour la nutrition humaine et la nutrition animale. Consommées sous forme de farine ou de grains entiers, elles fournissent 45 % des calories alimentaires à l'échelle mondiale. Le riz et le blé représentent plus de 80 % des céréales consommées dans l'alimentation humaine. La production de maïs est majoritairement destinée à l'alimentation animale (http://www.fao.org/faostat/fr/#search/céréales).

2.1.2 Description du panel

Outre leur importance agronomique, les céréales représentent un modèle de choix pour les études des mécanismes évolutifs et particulièrement de la polyploïdisation.

Le panel de céréales étudié dans ce travail compte huit espèces : le riz, *Brachypodium*, le blé tendre (hexaploïde), le blé dur (tétraploïde), l'orge, *Setaria*, le sorgho et le maïs. Sur le plan de la

systématique, ces 8 espèces appartiennent à 3 sous-familles des *Poaceae* qui comprennent la majorité des céréales cultivées : la sous-famille des *Panicoideae* incluant le maïs, *Setari*a et le sorgho, la sous-famille des *Pooideae* incluant les blés, l'orge, l'avoine et le seigle, et la sous-famille des *Ehrhartoideae* incluant le riz (Figure 28).

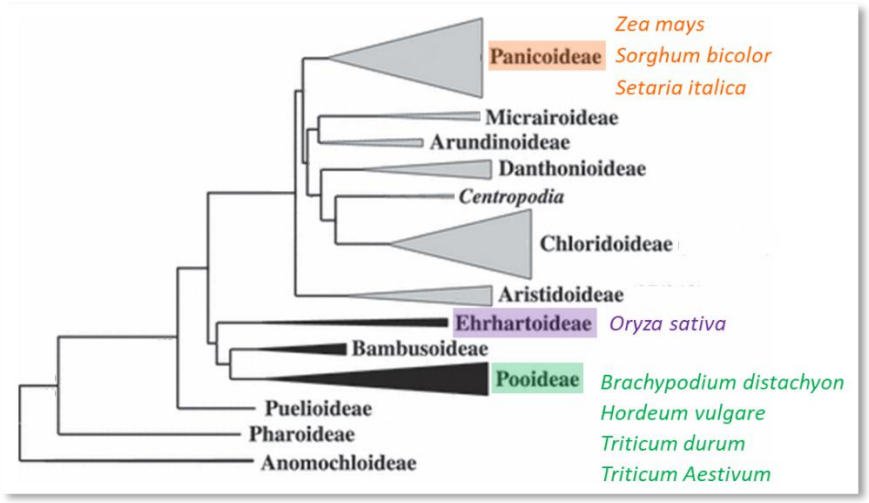
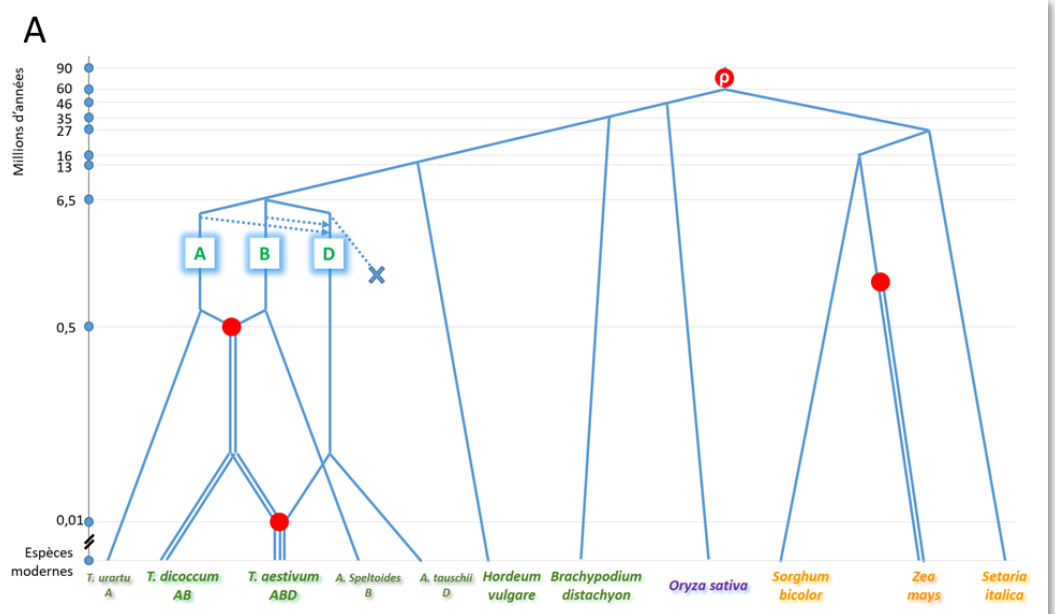


Figure 28 : phylogénie de Poaceae. Le clade BEP (Bambusoideae, Ehrhartoideae et Pooideae) est en noir, et PACMAD (Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae et Danthonioideae) est en gris. Les espèces du panel sont positionnées dans leurs sous-familles respectives. Adapté de Aliscioni et al., 2012.

Les 8 espèces descendent d'un ancêtre commun à 7 chromosomes, noté AGK7 pour *Ancestral Grass Karyotype*, qui a subi il y a 90 millions d'années une duplication totale de son génome pour atteindre une structure à 14 chromosomes, puis à 12 chromosomes (AGK12), suite à deux évènements de fusions chromosomiques. Toutes les céréales modernes dérivent de cet ancêtre à AGK à 12 chromosomes et ont donc toutes, du fait de la duplication ancestrale, notée ρ, un caractère paléotétraploïde (Murat *et al.*, 2017). Un premier évènement de spéciation survient il y a environ 65 millions d'années. Parmi les espèces qui en sont issues, seul le riz a conservé la structure ancestrale à 12 chromosomes et en ce sens il constitue l'espèce « pivot », image du génome AGK12. Les *Panicoideae* ont subi 2 fusions chromosomiques, il y a environ 27 millions d'années, formant un nouvel ancêtre intermédiaire à 10 chromosomes, dont le sorgho a conservé la structure. Le maïs a connu une nouvelle duplication totale de son génome, il y a environ 5 millions d'années, qui a donné lieu à un ancêtre intermédiaire à 20 chromosomes qui a subi plusieurs fusions chromosomiques pour aboutir à la structure moderne du maïs à 10 chromosomes. *Brachypodium* et les *Triticeae* se sont différenciés il y a environ 35 millions

d'années. Le génome de *Brachypodium* est le résultat de 7 événements de fusions chromosomiques aboutissant à la structure actuelle à 5 chromosomes. Les *Triticeae* ont acquis une structure ancestrale ATK, *Ancestral Triticeae Karyotype*, à 7 chromosomes *via* 5 fusions de chromosomes, il y a environ 26 millions d'années. L'orge et le blé ont divergé il y a environ 13 millions d'années, les deux espèces maintenant la structure à 7 chromosomes. Le blé a connu un nouvel évènement de spéciation, il y a environ 6,5 millions d'années, formant plusieurs espèces de blés diploïdes. Certaines de ces espèces diploïdes qui se sont successivement hybridées pour former le blé tétraploïde puis le blé hexaploïde, il y a respectivement environ 360 000 et 10 000 ans. Dans les deux cas il n'y a pas eu de remaniements structuraux majeurs, hormis une translocation, et par conséquent les sous génomes ont conservé leurs structures différenciées (Figure 29A).

A



B

Nom scientifique	Nom commun	Taille du génome (Mb)	% éléments répétés	Nombre de gènes annotés	Ploïdie
Setaria italica	Petit mil	515	46%	36k	diploïde (paléotétraploïde)
Sorghum bicolor	Sorgho	732	70%	28k	diploïde (paléotétraploïde)
Zea mays	Maïs	2100	85%	64k	tétraploïde
Oryza sativa	Riz	372	40%	42k	diploïde (paléotétraploïde)
Brachypodium distachyon	Brachypode à deux épis	270	28%	27k	diploïde (paléotétraploïde)
Hordeum vulgare	Orge	5 300	84%	36k	diploïde (paléotétraploïde)
Triticum dicoccum	Blé dur	10 400	80%	72k	tétraploïde
Triticum aestivum	Blé tendre	14 500	85%	108k	hexaploïde

Figure 29 : Volet A. Phylogénie des Poaceae. Les 4 WGD sont représentées par des points rouges. De l'évènement le plus ancien au plus récent : La WGD ancestrale commune à l'ensemble des espèces notée ρ, la duplication du génome maïs, l'hybridation des génomes de blés A et B formant le blé tétraploïde, l'hybridation du génome D au génome AB formant le blé hexaploïde. Les espèces modernes indiquées correspondent aux 8 espèces du panel complétées par les

blés diploïdes Triticum urartu, Aegilops speltoides et en Aegilops tauschii, plus proches représentants modernes des génomes A, B et D respectivement. L'échelle de temps est choisie pour représenter les interactions entre les génomes de blés. Les flèches pointillées illustrent les introgressions des génomes A et B au sein du sous-génome D. La croix représente l'absence de descendant direct du génome D originel. Adapté de Murat et al., 2017; El Baidouri et al., 2017. Volet B. Caractéristiques des structures génomiques des espèces du panel. Le nombre de gènes annotés est exprimé en milliers = k. Données issues des métadonnées des séquences génomiques considérées dans cette étude.

Les tailles des génomes des différentes espèces ont largement divergé sous l'effet de l'accumulation d'éléments répétés et, le cas échéant, suite aux WGD. Aux extrémités du spectre se positionnent *Brachypodium*, 270 Mb et 28% d'éléments répétés, et le blé hexaploïde, 14,5 Gb et 84% d'éléments répétés (Figure 29B).

2.1.3 Description des données omiques

Les données omiques utilisées pour mener les analyses sur lesquelles repose ce travail comprennent :

- **Les séquences annotées des génomes des 8 espèces du panel**. Les comparaisons des séquences des gènes orthologues et homéologues/dupliqués permettent de calculer les taux de **substitutions synonymes ks et non-synonymes ka**, et les longueurs de branches des arbres phylogénétiques construis.

- Les données de **Ka** et de **Ks**, complétée par des **données de diversité intraspécifique de type SNP** décrivant la variabilité nucléotidique des régions codantes du génome, pour le blé hexaploïde, *Brachypodium* et le maïs à partir de données de reséquençage à l'échelle populationnelle.

- **Données d'expression pour 3 espèces : le blé hexaploïde, le maïs et *Brachypodium***. Ces données correspondent à l'**expression des ARNm** dans le grain à trois stades morphologiquement bien définis de son remplissage, similaires pour les 3 espèces.

- **Données de méthylation pour deux espèces : le maïs et *Brachypodium*.** Ces données sont obtenues par séquençage bisulfite sur les mêmes échantillons définis précédemment.

L'analyse de ces données a permis de réaliser une étude des effets de la polyploïdie chez huit espèces de graminées. Les résultats obtenus ont fait l'objet d'un article présenté ci-après.

3 Article “Tracing 100 million years of grass genome evolutionary plasticity”

Cet article a été soumis à la revue *The Plant Journal*.

Sa version finale a été acceptée pour publication par *The Plant Journal*, le 24 février 2023.

La section « supplementary data » est présentée en annexes.

Tracing 100 million years of grass genome evolutionary plasticity

Running title: Evolution of genome organization and regulation during 100 million years of grass evolution

Arnaud Bellec[1]*, Mamadou Dia Sow[2]*, Caroline Pont[2], Peter Civan[2], Emile Mardoc[2], Wandrille Duchemin[2], David Armisen[2], Cécile Huneau[2], Johanne Thévenin[3], Vanessa Vernoud[4], Nathalie Depège-Fargeix[4], Laurent Maunas[4], Brigitte Escale[5], Bertrand Dubreucq[4], Peter Rogowsky[4], Hélène Bergès[1], Jerome Salse[2]**

[1]**INRAE/CNRGV US 1258.** 24 Chemin de Borde Rouge, 31320 Auzeville-Tolosane.

[2]**INRAE/UCA UMR 1095.** 5 Chemin de Beaulieu, 63100 Clermont Ferrand, France.

[3]**INRAE/AgroParisTech-UMR 1318.** Bat 2. Centre INRA de Versailles, route de Saint Cyr, 78026 Versailles CEDEX, France.

[4]**INRAE/CNRS/ENS/Univ. Lyon-UMR 879**, 46 allée d'Italie, 69364 Lyon Cedex 07, France.
[5]**Arvalis–Institut du végétal**. 21 chemin de Pau, 64121 Montardon, France.

*Contributed equally.
**Corresponding author.

Abstract
Grasses form a family of monocotyledonous plants that includes crops of major economic importance such as wheat, rice, sorghum and barley, sharing a common ancestor some 100 million years ago. The genomic attributes of plant evolution and adaptation remain obscure and the consequences of recurrent genome doubling (polyploidization), a major force in plant evolution, remain largely speculative. We conducted a comparative analysis of omics data from ten grass species to unveil structural (inversion, fusion, fissions, duplications, substitutions) and regulatory (expression and methylation) basis of genome plasticity, as possible attributes of plant long lasting evolution and adaptation.
The current study demonstrates that diverged polyploid lineages sharing a common WGD event often present the same patterns of structural changes and evolutionary dynamics, but these patterns are difficult to generalize across independent WGD events due to non-WGD factors such as selection and domestication of crops. Polyploidy is unequivocally linked to the evolutionary success of grasses during the past 100 My, although it remains difficult to attribute this success to particular genomic consequences of polyploidization, suggesting that polyploids harness the potential of genome duplication, at least partially, in lineage-specific ways. Overall, the current study clearly demonstrates that post-polyploidization reprogramming is more complex than traditionally reported in investigating one species and calls for a critical and comprehensive comparison across independently polyploidized lineages.

Introduction

Since Charles Darwin's seminal theory in 1859 (Darwin, 1859), the understanding of species evolution has relied on unveiling the forces promoting biodiversity through speciation and diversification, ultimately leading to morphological and phenotypic innovations. In this respect, Susumu Ohno (S Ohno, 1970) proposed polyploidy – also referred to as whole genome duplication (WGD) – as a key contributor to such innovation. Polyploidy is a condition of having multiple (>2) sets of chromosomes that coexist in one nucleus and can be stably inherited by progenies. Polyploids are formed in two ways, autopolyploidization (genome doubling involving the same parental species) and allopolyploidization (genome doubling involving genomes from two parental species), (Leitch and Bennett, 1997). A further distinction is related to the timing of this event. Paleopolyploids are species that experienced polyploidization in their ancient past and their genomes have been subsequently re-diploidized through structural and functional reorganization. Neopolyploids are species that experienced polyploidization more recently and still possess distinguishable sets of parental chromosomes.

Polyploid species are widespread across the animal kingdom, including frog (Schmid *et al.*, 2015), fish (Taylor *et al.*, 2003; LeComber and Smith, 2004; Dehal and Boore, 2005; X., L., Liu *et al.*, 2017), insects (Li *et al.*, 2018), and mammals (Acharya and Ghosh, 2016). In plants, polyploidy is ubiquitous in flowering plants (Van De Peer *et al.*, 2017) and widely found in lower plants, such as gymnosperms (Z., Li *et al.*, 2015), ferns (Wood *et al.*, 2009; Hidalgo *et al.*, 2017), and diatoms (Parks *et al.*, 2018). It has been proposed that polyploidy is implicated in the generation of phenotypic diversity (Soltis and Soltis, 2009; Landis *et al.*, 2018; Leebens-Mack *et al.*, 2019), species diversification (Levin and Soltis, 2018), crop domestication (Salman-Minkov *et al.*, 2016), and adaptation (Vanneste, Baele, *et al.*, 2014), all of which is attributed to the enhanced genomic plasticity generated by WGD (Leitch and Leitch, 2008). Thus, numerous studies have suggested that polyploidy contributed to the evolutionary success of extant angiosperm species, including the model plant *Arabidopsis* (Thomas *et al.*, 2006) and major crops like sorghum (Paterson *et al.*, 2009), maize (Schnable *et al.*, 2011), *Brassicaceae* (Cheng *et al.*, 2012; Murat, Louis, *et al.*, 2015), wheat (Pont *et al.*, 2013), cotton (Renny-Byfield *et al.*, 2015; Paterson *et al.*, 2012) and soybean (Schmutz *et al.*, 2010).

However, several fundamental questions about polyploidization remain open. What are the structural and functional features of the new balance between genomic plasticity, stability and evolvability? To what extent does polyploidization enhance molecular evolution, epigenetic changes and alterations in gene expression in duplicated genomic regions and genes? And what mechanisms establish these changes? Addressing these questions requires a reconstruction of the gene content and genome organization of extinct ancestors predating the speciation and WGD events. The reconstruction of the ancestral genomes allows precise identification of major karyotype changes and associate them with evolutionary consequences (Salse 2012). In this study, we make a wider use of this strategy, expanding it beyond a single species and WGD, towards a series of speciation and polyploidization events, covering several species of the grass family over the last 100 million years (my). The objective is to provide generalized description of the structural and functional consequences accompanying polyploidization events during plant evolution.

Results

Consequences of polyploidization events on karyotype remodeling

Grasses (*Poaceae* family) went through a whole genome duplication event around 100 my ago (referenced as ρ), *i.e.* before the divergence of the *Ehrhartoideae* (including rice), *Pooideae* (including *Brachypodium* as well as the *Triticeae)* and the *Panicoideae* (including sorghum, setaria and maize*)* subfamilies*.* After the ancestral shared ρ tetraploidization, additional lineage-specific polyploidization events occurred in wheat (tetraploidization and hexaploidization, 0.36 and 0.01 My ago, respectively) and maize (tetraploidization 5 My ago). Grasses can therefore be considered as the ideal angiosperm model to investigate the fate of both paleo- and neo-polyploidy events on genome organization, regulation and ultimately adaptation of the resulting species. Comparative genomics, aiming to reveal the fate of shared as well as lineage-specific genes, was conducted through the reconstruction of ancestral *Poaceae* genomes. The ancestral genome is a 'median' or 'intermediate' genome consisting of the most parsimonious gene order, based on the extant genes conserved between the modern species investigated (Murat et al. 2017). A pivot species was chosen for each of the investigated subfamilies on the basis of the smallest numbers of historical polyploidization events (*i.e.* rice, *Brachypodium,* sorghum). The subsequent comparative analyses used these pivots as a reference. Ancestral grass karyotype (AGK) were inferred by integrating paralogies and orthologies (following a synteny-based approach) from the eight investigated genomes (rice, *Brachypodium*, barley, tetraploid and hexaploid wheat, setaria, maize sorghum), defining independent contiguous ancestral regions (CARs, also referred to as protochromosomes). This resulted in a pre-ρ AGK of 7 protochromosomes (hereafter AGK7) with 10,286 ordered protogenes and a post-ρ AGK of 12 protochromosomes (hereafter AGK12) with 16,560 ordered protogenes (Figure 1a, Supplementary Figure 1a), refining and enriching (above 10%) the ancestral gene content published previously (Pont et al. 2019). Protogenes are enriched in gene ontology (GO) terms related to basic cellular functions such as transferase, transporter, signaling activity (p-value < 0.05, Supplementary Figure 2a). The comparative genomics data produced here are publicly available at https://urgi.versailles.inra.fr/synteny/ (Supplementary Data Set 1).

Comparison of the gene order between the inferred ancestors (AGK7 and AGK12) and the eight modern genomes allowed to identify losses of collinearity (gene-to-gene adjacency), defining synteny breakpoints (SBPs, Supplementary Figures 1 and 3, Supplementary Table 1). SBPs are derived from ancestral chromosome fusions, fissions and translocations in the course of evolution. We propose that AGK7 was duplicated (ρ paleotetraploidization) to reach AGK12 intermediate through fissions (2) and translocations (2), overall defining 6 SBPs. Remarkably, the synteny of AGK12 and rice genomes is fully conserved, while 4 (2 fusions), 9 (3 fusions and 1 translocation) and 38 (18 fusions) SBPs since AGK12 have been identified in sorghum, setaria and maize, respectively. By comparing barley to diploid, tetraploid and hexaploid wheat genomes, we reconstructed the ancestral *Triticeae* karyotype (ATK7) made of seven protochromosomes covered by 12,718 conserved genes. Relative to AGK12, ATK underwent 5 fusions and a complex interplay of translocation shaping the modern *Triticeae* chromosomes 4 and 5, leading to 10 SBPs. The wheat A subgenome (in tetraploids and hexaploid) shows an additional SBP corresponding to the 4A/7B translocation, leading to a total of 11 SBPs. Following the karyotype evolution from the basis of the modern grasses and the basis of the *Triticeae* (AGK12 and ATK7, respectively), the

numbers of SBPs leading to the modern genome structure greatly differ between lineages, without any shuffling event or SBP shared among the eight investigated species.

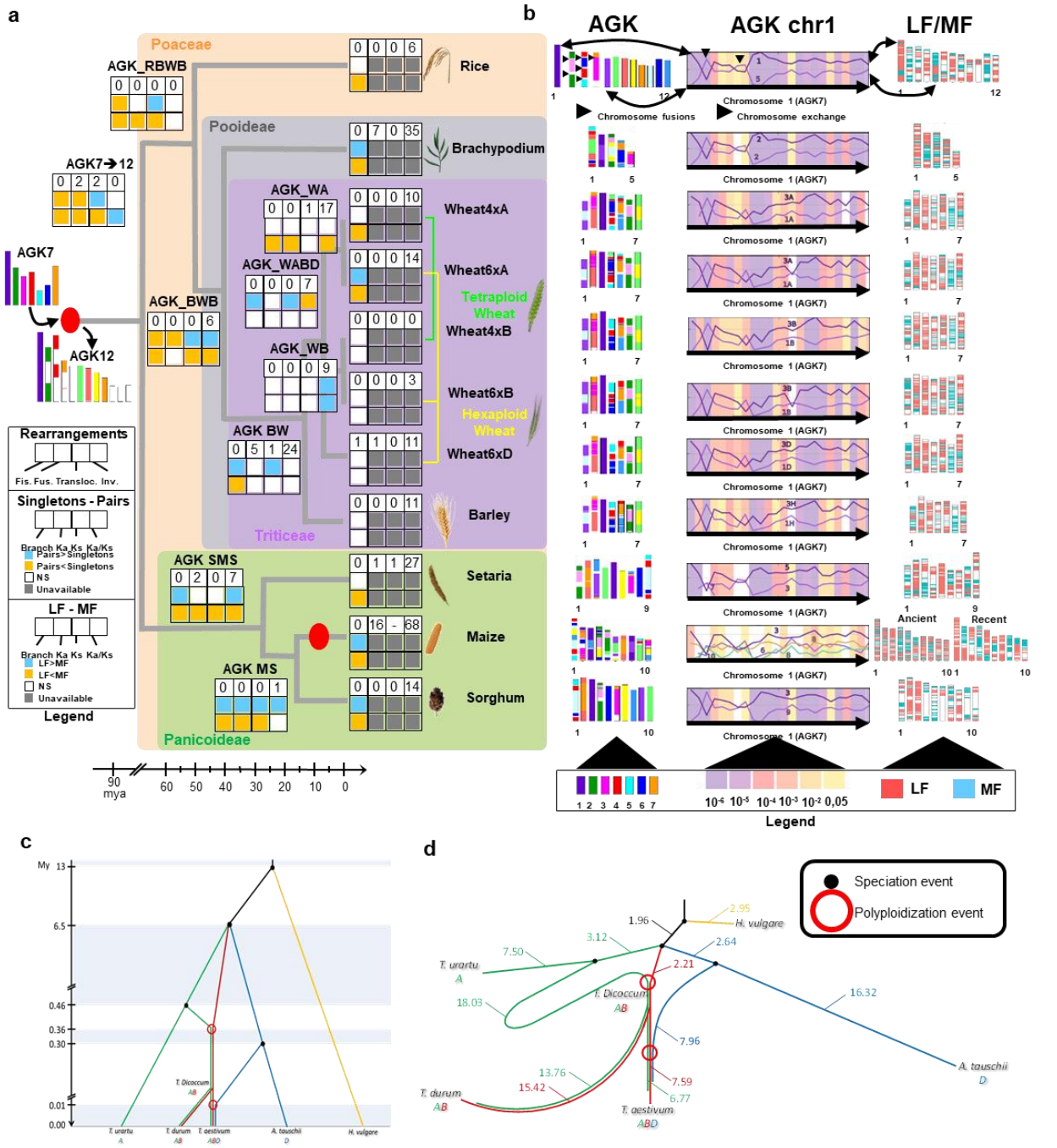


Figure 1: Evolution of grass genome structures. a. A phylogenetic tree of grass evolution according to the time scale (below) in My. Common ancestors of extant species (at the right) reported at each tree node are abbreviated according to the initial letters of the descending species, with e.g. BWB ancestor corresponding to the speciation time point of Brachypodium-Wheat-Barley. For each ancestor, the number of chromosomal fissions (fis.), fusions (fus.), translocation (trans.) and inversions (inv.) are shown in the top rows of the mini-tables, together with colour-coded differences in branch length, Ka, Ks, Ka/Ks between singletons and pairs (middle row) and LF and MF

genomic fractions (bottom row; see legend at the bottom left). The ancestral WGD and maize specific WGD are indicated by red ovals. b. Left panel: extant genomes are illustrated with a colour code for the 7 ancestral chromosomes (AGK7 color code). Central panel: Graphs show the retention of ancestral genes for each pair of the duplicated blocks (ρ WGD) corresponding to the ancestral chromosome 1. Significant differences of gene retention between the duplicated blocks are indicated by a colour code (see legend below). For maize, both the ρ WGD and the recent maize-specific WGD is considered (four duplicated blocks instead of two). Right panel: modern genomes are illustrated with the LF (Least Fractionated) and MF (Most Fractionated) genomics compartments (red for LF-genes and blue for MF-genes). c. Phylogenetic tree of Triticeae scaled according to estimated time (in My) of speciation (black dots) and polyploidization (red circles) events d. Variations of substitution rates in the Triticeae lineage. The lengths of the branches are proportional to the substitution rates expressed in number of substitutions per site per billion years, highlighting differences of molecular evolution between species and subgenomes.

In addition to the large chromosomal rearrangements, we investigated inversions during grass evolution (Figure 1a, Supplementary Table 2, Supplementary Figure 4a). Overall, the fraction of the genome covered by inversions ranges from 0.2% in rice to 22.8% in the tetraploid wheat subgenome B (Supplementary Figure 4b). Inversion events have particularly impacted *Triticeae* genomes, covering 17.4% of the barley genome, 19.4% of the hexaploid wheat subgenome D and 20.6–22.8% of the wheat subgenomes A and B in the tetraploid and hexaploid contexts. Wheat subgenome A, both in the tetraploid (64 inversions since AGK12) and hexaploid (68 inversions) contexts, appears more dynamic than the B (46 and 49 inversions in tetraploid and hexaploid wheat, respectively) and D (48 inversions) subgenomes. Noteworthy, most of the subgenome A extra-inversions occurred after the tetraploidization event (Supplementary Figure 4c). The maize genome went through an increase in inversions (68) after its divergence with sorghum (14 inversions) and the polyploidization event dating back to 5 My ago. Although some of the inversions observed in maize could pre-date WGD and hence be doubled by it, the several-fold difference in the number of inversions and the genomic space covered by them (14.1% in maize *versus* 12.5% in sorghum) clearly indicate the role of polyploidizations in promoting genome shuffling. Inversions appear to be higher gene density comparing to 1000 simulations, by positioning the inversions randomly on the chromosomes of the investigated species (Supplementary Figure 4d), and tend to be located in gene rich High Recombination (HR) genomic regions compared to gene poor Low Recombination (LR) regions (Supplementary Figure 4e). We have also observed signs of TE involvement in genomic rearrangements, particularly on the chromosome 6A of hexaploid (cv. Chinese Spring) and tetraploid wheat (wild emmer Zavitan and durum cv. Svevo) showing the highest rate of inversions (Supplementary Figure 5a-c). We found an insertion of an LTR-retrotransposon Jorge (*Copia* superfamily) at the immediate boundary of an inversion detected between Chinese Spring and Zavitan, and similarly, an insertion of Laura (*Gypsy* superfamily) bordering on an inversion detected between Chinese Spring and Svevo.

Impact of polyploidization events on sequence divergence

To study gene divergence following polyploidization, we generated 15,810 gene trees from a total of 304,549 genes representing 68% of the genes annotated in the eight investigated species. Out of the 15,810 gene trees, 2,599 and 3,851 correspond to AGK7 (pre-ρ) and AGK12 (post-ρ) genes, respectively, conserved (*i.e*. present with at least one homolog) in all investigated species. Phylogenetic tree reconciliation lead to the identification of 22,334 ancestral genes at the time of the maize WGD, 30,567 ancestral genes at the speciation of the AGK12 ancestor (complementing

the catalog of 16,560 protogenes from the synteny-based approach described in the previous section), and 15,435 ancestral genes at each node of the speciation of the investigated species (Supplementary Figure 6a, Supplementary Table 4). Gene phylogenies allowed us to investigate substitution rate changes in conserved genes inherited from a particular speciation or duplication event, expressed as substitutions per site and per billion years. Both the ancient WGD (ρ) and the recent maize WGD are associated with an increased rate of substitutions (4.11 and 9.79, respectively) as well as speciation events (5.04 and 4.30 at the basis of the *Pooideae* and *Panicoideae*, respectively) compared to all the closest relatives (<3.23). In the case of the ancient (ρ) WGD event, the substitution rate has undergone a net slowdown in both lineages resulting from the speciation of AGK12, with 2.15 for the *Pooideae* and 2.04 for the *Panicoideae*. In both sub-families, the internodes leading to speciation (BWB ancestor, BW ancestor and MS ancestor) appear to evolve faster (with rates of 5.04, 3.23 and 4.3 respectively) compared to post-speciation tree branches (with 2.74 for rice, 2.17 for *Brachypodium*, 3.07 for barley, 2.84 for wheat, 2.31 for *Setaria*, 1.93 for sorghum), Supplementary Figure 6b and Supplementary Table 5.

In order to gain better insights into the consequences of recent polyploidization on substitution rates, we focused on the complex *Triticeae* lineage represented by 7 ATK protochomosomes covered by 12,718 protogenes, in selecting a set of 3,905 single-copy orthologs shared by barley, diploid (*Triticum urartu* and *Aegilops tauschii*), tetraploid and hexaploid wheats (Figure 1c). Considering a shorter time scale in *Triticeae* evolution (13 My, Supplementary Figure 6), the count of substitutions are lower than those previously described in grasses (Supplementary Table 5). Nevertheless, Such *Triticeae* specific gene set provides a better understanding of the recent divergence between wheats subgenomes, contrasting the speed of the different lineages in response to polyploidization (Figure 1d, Supplementary Figure 6c, Supplementary Table 5). Barley and the wheat lineage prior to the speciation and polyploidization events display low substitution rates, ranging from 1.96 to 2.95 (substitutions per site per billion years). The highest substitution rate of 18.03 was observed for the A lineage leading to polyploid wheat, relating to the period after the divergence from *T. uratu* (0.46 My ago) and before the hexaploidization event marked by the divergence of the A subgenomes in the tetra- and hexaploid contexts (0.01 My ago). This long internode spans the tetraploidization event, which, however, cannot be placed exactly. When focusing on the terminal branches, there are two apparently discontinuous categories of substitution rates. Low substitution rates (6.77–7.96) were obtained for the A, B and D subgenomes in the hexaploid context, and the diploid A lineage (*T. urartu*). High substitution rates (13.79–16.32) were obtained for the A and B subgenomes in the tetraploid context, and the diploid D lineage (*Ae. tauschii*). While the observed branch lengths indicate that substitution rates could have been accelerated in the tetraploids or alternatively decelerated in the hexaploids, no general pattern of substitution rate change in relation to polyploidizations can be concluded. Associating substitution rates to polyploidization events is further complicated by the wide range observed at the diploid level (7.5 in *T. urartu*; 16.32 in *Ae. tauschii*), and the uncertain effects of artificial selection (purifying selection; the cost of domestication) that accompanied the domestication of tetra- and hexaploid wheats.

Impact of polyploidization events on gene content

Subgenome dominance, *i.e.* a preferential retention of genes in one of the two homoeologous genomic regions, is another presumed consequence of polyploidization. To investigate this biased fractionation of the duplicated regions, we developed a window-based statistical approach defining the Least Fractionated (LF), Most Fractionated (MF), and unbiased compartments of the genome (Figure 1b, Supplementary Figure 7a). Among the 447,744 genes annotated in the eight investigated species, a total of 158,124 (35% of annotated genes) and 74,311 (17% of annotated genes) were classified as LF or MF respectively. The LF compartment contains genes enriched in functions related to signaling processes (transducer, transferase, transporter, kinase, etc.) while the MF compartment is enriched in transcription activity (binding, transcription, etc.), Supplementary Figure 2b. A similar approach has been applied for the recent duplication in maize, resulting in the annotation of 29,148 (46% of annotated genes) and 17,455 (27%) as LF- and MF-located genes, respectively (Supplementary Figure 7b). Overall, we deliver a statistically-based assessment of 187,272 LF- and 91,766 MF-located extant genes in respect to the ancestral shared ρ and maize-specific WGDs (Supplementary Data Set 1 and Supplementary Table 4). For each of the pre-duplication chromosome, the two post-polyploidization (*i.e*. homoeologous) blocks are identified as either LF, MF or unbiased. Fraction (LF becoming MF or *vice versa*) changes on a duplicated block deriving from a polyploidization event may then reflect homoeologous chromosome exchanges. Based on this assumption, it appears that each of the post-WGD chromosome pairs, derived from AGK7 following the ancestral shared ρ events, experienced homoeologous chromosome exchanges particularly in the (peri-)centromeric region (Figure 1b, Supplementary Figure 7).

Complementary to our analysis of genes located in LF and MF compartments, we investigated evolutionary trajectories of genes that are retained in a single copy (singletons) or two copies (Pairs) after WGD (Supplementary Table 4). From the gene-based phylogenetic approach described above (15,810 gene trees), the singletons were defined as genes with an ancestral copy in AGK7 retained in a single copy in the eight extant species, thus defining 156,560 singletons (35% of annotated genes) in the extant genomes. The same strategy was followed in defining 15,157 singletons (24% of annotated genes) following the maize-specific WGD from AGK12. Conversely, ancestral genes in AGK7 retained in two copies in the extant species defined 113,837 genes as pairs (25% of annotated genes), while 14,293 genes (23% of annotated genes) were identified as pairs in maize when considering the maize-specific WGD and AGK12. Complementing the previous catalog of singletons and pairs based on phylogeny, singletons and pairs were also inferred using the synteny-based approach, delivering 16,560 ordered protogenes from AGK12 (post-ρ ancestor) and 10,286 ordered protogenes from AGK7 (pre-ρ ancestor). This led to the identification of 19,032 genes as pairs and 75,662 as singletons. By definition, the LF-compartment is expected to contain more singletons relative to the MF-compartment, while genes in the MF-compartment are more likely to be in the paired category (each gene loss in the MF-compartment produces a singleton in the LF-compartment; scarcity of gene losses in the LF-compartment leads to a lack of singletons in the MF-compartment). In accordance with this expectation, both inferences of singletons and pairs (phylogeny- or synteny-based approaches) identified sets of GO terms similar to the ones obtained for LF- and MF-compartments, respectively. Singletons are enriched in functions related to signaling processes, with the most

significant GO terms for transferase, transporter, hydrolase, kinase (Supplementary Figure 2c). Gene pairs are enriched in gene functions related to transcription activity with top significance for binding and transcription (FDR < 0.05 using rice as reference, Supplementary Figure 2c).

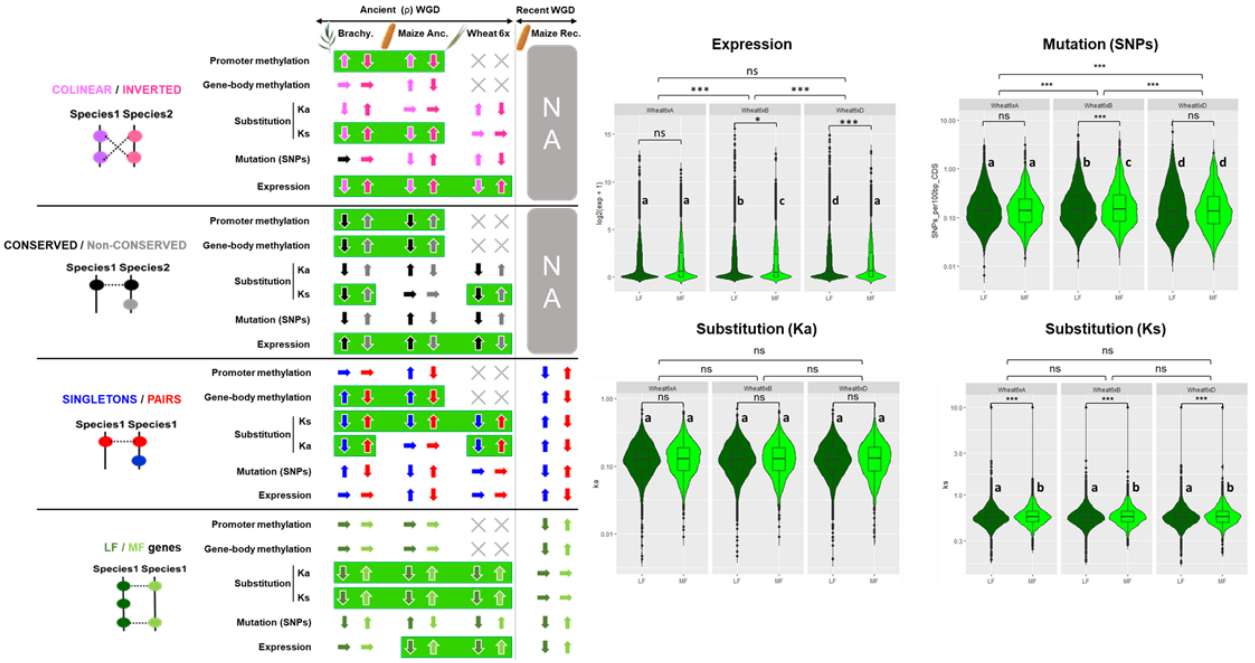
Regarding the sequence divergence of genes (Figure 1a, Supplementary Table 6), no difference in non-synonymous substitution rates (Ka) has been observed between singletons and pairs in 6 out of 9 investigated MRCAs. However, regarding synonymous substitutions (Ks) rate, genes in pairs displayed a significantly higher rate compared to singletons for 6 out of the 9 investigated MRCAs. These differences were partially reflected in the ratios of non-synonymous to synonymous substitutions (Ka/Ks), with higher values observed in pairs for 4 out of 9 investigated MRCAs, except for the wheat ABD ancestor exhibiting a higher Ka/Ks ratio for singletons. Differences in Ka, Ks and Ka/Ks between singletons and pairs are not significant for the remaining ancestors. To complement the investigation of grass gene sequence evolution, we examined branch lengths on phylogenetic trees (Supplementary table 6). In 4 out 6 of the more recent MRCAs (SMS ancestor (27 My), MS ancestor (16 My), BW ancestor (13 My), WABD ancestor (6.5 My)), as well as 4 out of 11 extant species, pairs have evolved faster than singletons. This is, however, reversed in the older MRCAs (AGK12 (60 My), RBWB ancestor (46 My) and the BWB ancestor (35 My)) where singletons have evolved faster than pairs. Taken together, some of these observations support the notion that sequence divergence following WGD depends on whether the gene is conserved in multiple copies (pairs) or not (singletons). When observed significant for 58% (21 among 36) of the Ka-Ks-KA/Ks values obtained for the 9 inferred ancestors, not all of the significant differences go in the same direction with pairs evolving faster than singletons in 71% (15 among the 21) significant differences.

At post-polyploidization block level, comparisons between genes located in LF and MF fractions ((Figure 1a, Supplementary Table 6) revealed that LF-genes exhibit significantly lower molecular evolution rates than MF genes in five out of nine MRCAs, considering both synonymous and non-synonymous substitution rates (LF/MF ratios ranging from 1.017 to 1.072 and from 1.026 to 1.189 for the synonymous and non-synonymous substitutions, respectively). Moreover, ancestors where the differences between the LF and MF compartments were not significant are the more recent ancestors (*i.e.,* more distant from the original ρ WGD event) suggesting that the polyploidization effect on gene sequence is no longer active after such a long period of time or that allopolyploidization (wheat hexaploidization) has less effect than autopolyploidization (ancestral ρ WGD) Ratios of non-synonymous to synonymous substitutions (Ka/Ks) were found to have a more complex pattern. The Ka/Ks ratio is lower for LF-genes than for MF-genes in 3 out of 9 ancestors (BWB, SMS and WA ancestors) and higher in AGK12 and WB ancestors.

In addition to Ka and Ks dynamics, the analysis of branch lengths displayed a pattern coherent with previous results, but with a finer temporal inference (Figure 1a, Supplementary Table 6). The genes in the MF fraction displayed significantly longer branches compared to the LF-located genes in 7 out of the 9 MRCAs and 7 out of the 11 extant species. These results indicate that sequence divergence of genes after polyploidization is affected by whether the gene belongs to the LF or MF fraction of the genome, with LF-located genes never evolving faster with statistical significance. However, this phenomenon of differential evolution rates appears to have occurred soon after the polyploidization event, with waning effect over longer periods of time.

Impact of polyploidization events on genome regulation (expression, methylation).

The evidence presented above (Fig. 1b) documents the extent of biased gene retention after WGD, *i.e.* the subgenome dominance in grasses. Perhaps the most intriguing question pertains to the drivers and mechanisms of this phenomenon. Recently, evidence of a possible involvement of repetitive elements reactivated *via* modification of small RNAs and epigenetic marks has been presented (Wang et al. 2022). In order to test this hypothesis, we generated *Brachypodium*, maize and wheat RNAseq data and whole genome bisulfite sequencing data, and collected publicly available SNPs dataset (Supplementary Data set 2, Supplementary Table 7). We investigated the relationship between the genome fractionation status (genes located in inverted or collinear blocks, LF or MF blocks, conserved *vs.* species-specific genes, pairs *vs.* singletons; Fig. 1a), evolutionary dynamics (substitution rates and nucleotide diversity) and regulation (gene expression, DNA methylation). Conclusions are raised below when omic variations in respect to the ancestral shared ρ WGD are consistent between at least two species (maize, wheat, *Brachypodium*) without a conflict (*i.e*. with no significant signal) in the third species (Figure 2 a-d).



<u>Figure 2</u>: Regulation dynamics during grass evolution. a. Omics differences such as methylation (promoter and gene-body), Substitution (Ka and Ks), polymorphisms (SNPs) and expression are illustrated with arrows (↑ for significant increase; ↓ for significant decrease; -> for no significant difference). The comparisons were performed between: (a) collinear *vs.* inverted, (b) conserved *vs.* non-Conserved, (c) Singleton *vs.* pair and (d) LF- *vs.* MF-located genes. For each species and for each of the omics data considered the first arrow (left) corresponds to collinear-conserved-singleton-LF genes and the second arrow (right) corresponds to inverted-nonconserved-pair-MF genes (on the a, b, c and d panels, respectively). Omic variations in respect to the ancestral shared ρ WGD that are consistent between at least two species without a conflict (i.e. with no significant signal in the third species) are highlighted with green background. e. Omics variation (expression, mutations, substitution) between wheat post-polyploidization compartments (LF *vs.* MF) inherited from the ancestral shared ρ WGD in A-B-D subgenomes regarding the wheat lineage-specific allopolyploidization.

We first assessed the possible role of genomic inversions in modification of gene regulation, following the idea that local changes in gene order (synteny decay) can disrupt cis-trans-regulatory interactions. Genes in inverted regions indeed exhibit, higher expression and lower promoter methylation levels, in addition to higher non-synonymous substitutions rate (Figure 2a, Supplementary Figure 8). We also found lower expression and higher methylation levels (at promoter and gene body), as well as higher synonymous substitution rates in species-specific genes compared to the conserved ones (Figure 2b). Among the conserved genes, ancient pairs exhibit lower gene-body methylation level and elevated levels of molecular evolution (ka, ks) compared to singletons (Figure 2c). The recent maize duplication displayed distinct patterns, with only the gene-body methylation mimicking the omics variation observed for the ancestral ρ duplication. When comparing duplicated blocks, genes located in the MF fraction are more expressed in all species except *Brachypodium* and exhibit higher Ka/ks rates (for all species except the recent maize WGD). When focusing on the most extremely fractionated LF and MF blocks at the whole genome level, defining fast evolving (FE, MF regions with the highest loss of ancestral genes) and slow evolving (SE, regions with the highest retention of ancestral genes) regions, the same trends were observed for gene expression, Ka, Ks and SNPs (Supplementary Figure 8b). However, in maize, FE regions inherited from the ancestral ρ duplication were more methylated than SE, while no methylation bias was observed between FE and SE regions inherited from the maize-specific duplication (except for CG promoter), in contrast to what is observed between the LF-MF compartment from the recent maize duplication.

To better understand the effect of recent polyploidizations, we first investigated divergence between wheats subgenomes (A, B and D). Subgenomes display divergence in expression (A>D>B) and diversity (B>A>D) with no difference in substitution rates (Ka and Ks), (Supplementary Figure 9). When ancient blocks (LF and MF) inherited from the ancestral (ρ) duplication are considered for each of the subgenomes (A, B and D), defining six genomic compartments in wheat genomes (LF and MF for each A, B and D), the LF-D accumulates less polymorphisms (SNPs) than in A and B (LF-B > LF-A > LF-D), being the most stable compartment of the wheat genome. The same patterns of SNP differences were observed for MF-genes with MF-B > MF-A > MF-D. Expression data show a more complex profile, where the B subgenome is the less expressed in both LF and MF compartments, the A subgenome appeared to be the most expressed in the LF compartment while the A and D subgemones are the most expressed in the MF compartments (Figure 2b, Supplementary Table 8). The same analysis was performed at the duplicated genes level (Singletons *vs.* Pairs). No significant differences were observed for expression, synonymous and non-synonymous substitutions between the A, B and D subgenomes. However, the D subgenome accumulated less polymorphisms for both singletons and pairs compared to the A and B subgenomes (Supplementary Figure 10) illustrating that additional processes, such as domestication and selection, have shaped the wheat genome over other genomic consequence in response to polyploidizations either recent (between A, B and D) or ancient (between LF and MF).

Impact of polyploidization events on conserved and duplicated genes

We then investigated regulation (expression and methylation) differences between pairs of conserved (between species) or duplicated (within species) genes (Figure 3). To do so, we first

defined profiles for each gene by concatenating their expression or methylation status (0 for non-expressed or non-methylated, 1 otherwise) for three comparable developmental stages (stage 1, 2 and 3) of the grain development in maize, *Brachypodium* and wheat. From a repertoire of 1997 genes showing a perfect 1-2-3 gene-to-gene copy relationship in *Brachypodium*, maize and wheat, respectively (referred to as ohnologs), we found that half of the ancestral genes conserved the same expression profile in all three species (43% are expressed in all species and 6% are not expressed in all three species, Figure 3a). When looking at duplicated genes in the same species, we observed that 70% and 75% of the duplicates share the same expression profile in maize and hexaploid wheat, respectively, while 13% (249) and 5% (85) of the duplicates showed 'On/Off' expression pattern (*i.e.* one copy of the pair expressed, the other copy being silenced). Between the three species, a set of ancestral ohnologous genes with On/Off pattern have been identified, corresponding to 10%, 9% and 1% genes conserved between *Brachypodium*/maize, *Brachypodium*/wheat and maize/wheat respectively (Figure 3a, Supplementary Table 9).

Similarly, a repertoire of 2,953 genes showing a perfect 1-2 gene-to-gene relationship between *Brachypodium* and maize was used to assess DNA methylation differences among conserved genes (methylation cut-off = 5 CpG at 10X of coverage, Figure 3a, Supplementary Table 9). Methylome analysis shows stronger differences between the three developmental stages in *Brachypodium,* where 46% of conserved genes show distinct pattern between stages, probably due to major DNA methylation variations between the two species during early embryonic stages of grain development. In maize, 34% of ohnologs show methylation differences between the three developmental stages. Comparison between *Brachypodium* and maize ohnologs indicates higher methylation level in maize genes (22% with 1-1-1 and 23% with 0-0-0 between maize duplicates) than in *Brachypodium* (5% with 1-1-1 and 49% with 0-0-0 profile). We found 16% of the ancestral genes retaining the same methylation patterns between the two species with 14% corresponding to 0-0-0 profile (unmethylated genes in the three stages in both species). In maize, this percentage increase to 45% with 22% of duplicates corresponding to 1-1-1 and 23% to 0-0-0 expression profiles. We then tested whether these differences in gene expression and DNA methylation between the ohnologs and duplicates (in maize) could be related to the presence of transposable elements (TEs) in proximity of genes. No significant correlation has been observed between the presence of TEs in the vicinity of genes and either the differences in expression or methylation levels between conserved or duplicate genes (Figure 3b, Supplementary Figure 11).

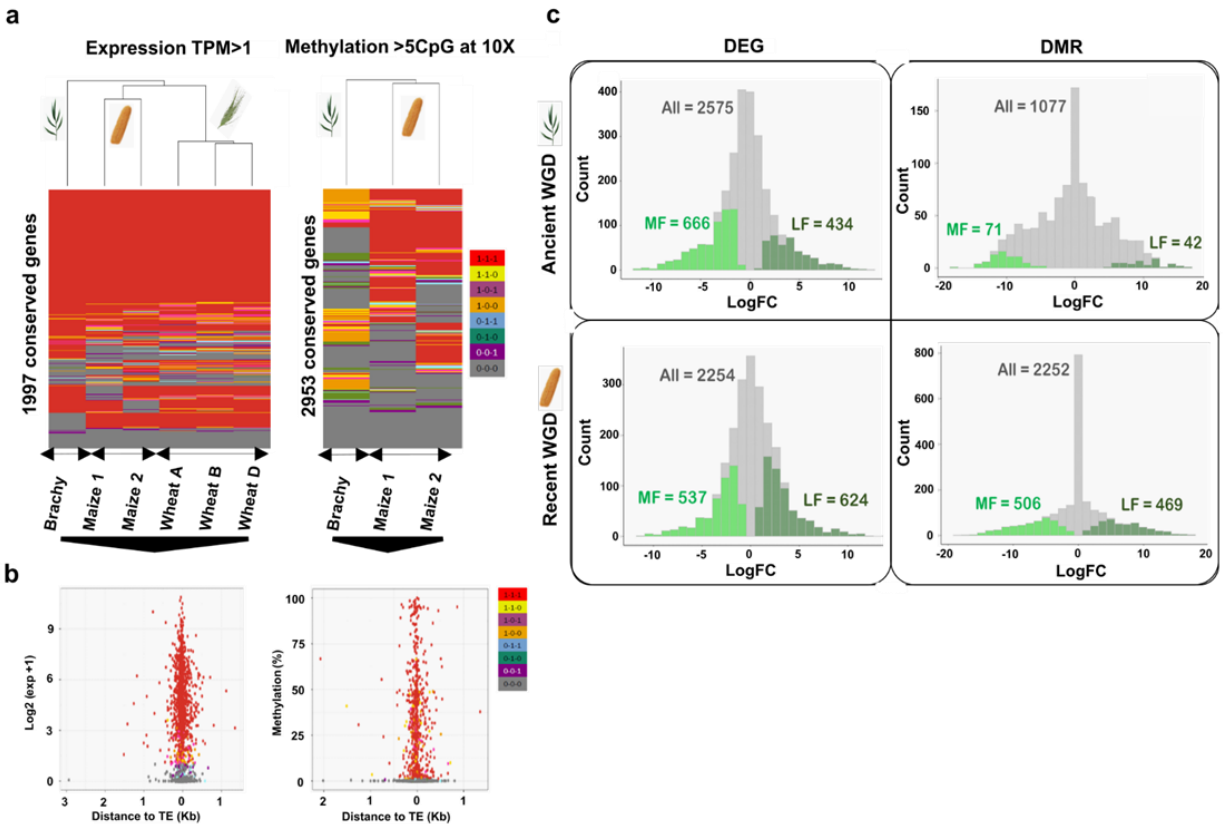


Figure 3: Regulation of conserved and duplicated genes. a. Omics variation between 1,997 genes showing a perfect 1-2-3 gene-to-gene relationship between *Brachypodium*, maize and wheat (for expression at the left) and 2,953 genes showing a perfect 1-2 gene-to-gene relationship between *Brachypodium* and maize (for methylation at the right) using a color code illustrating expression/methylation profiles of the three developmental stages with 0 corresponding to no detected expression/methylation and 1 corresponding to detected expression/methylation. I.e. profile 1-1-1 (red) indicates genes expressed/methylated in the three developmental stages considered. b. Average distance of the closest TE for duplicated pairs showing differences in expression and methylation in maize. The x axis represents the distance to the nearest transposable element (TE) from the transcription start site (TSS), and the y axis indicates the gene expression (in logarithmic scale, left panel) or DNA methylation (in %, right panel) levels, respectively. c. Differences in expression (left, with DEGs for differentially expressed genes) and methylation (right, with DMGs for differentially methylated genes) between gene pairs in *Brachypodium*, maize and wheat deriving from the ancestral shared ρ WGD (top) and for the maize-specific duplication (bottom). Pairs that are differentially expressed (FDR < 0.05) and differentially methylated (FDR < 0.05) in LF compared to MF are shown in dark green (upregulated LF-genes) and light green (downregulated MF-genes)

We then investigated the expression fate of duplicated genes after WGD in *Brachypodium*, maize and wheat. After assigning the two copies of duplicates to their corresponding LF or MF compartments, we analyzed their expression and methylation differences (Differentially Expressed Genes, *i.e.* DEGs and Differentially Methylated Genes, *i.e.* DMGs, Figure 3c). Most previous studies focus on expression differences between duplicated genes, with one copy expressed while the other one remaining silenced. However, only about half of the duplicated genes we investigated (for both ancient ρ and recent maize WGDs) exhibit expression differences, with the remaining pairs having the same expression pattern. When expression divergence is

observed between pairs, genes from the MF compartment are overexpressed for ancient WGD (666 vs. 434 genes in the LF compartment, Chi-square p-value=$2.65e^{-12}$) while genes in the MF compartment are significantly less expressed in the case of the recent maize WGD (537 vs. 624 genes, Chi-square p-value=$1.07e^{-2}$). DNA methylation also displays a distinct pattern between pairs depending on the WGD events investigated (ancient *vs.* recent). The vast majority (90%) of ancient duplicates do not exhibit any DNA methylation differences, likely due to the age of the ρ WGD (100 My), which is expected to confound changes established shortly after WGD. Nevertheless, significantly higher number of observed DMGs are hypomethylated in the MF compartment (71 in MF and 42 in LF, Chi-square p-value=$6.37e^{-3}$). The recent duplication event in Maize (5 My ago) offers more resolution, with about half of duplicated genes (43%) exhibiting DNA methylation differences. Similar numbers of these DMGs are hypomethylated in MF and LF (506 and 469, respectively; Chi-square p-value=0.236). Finally, duplicated pairs from the MF compartment exhibit higher synonymous and non-synonymous substitution rates for both the ancient and recent maize-specific WGDs (Supplementary Figure 12).

Interplay between genomic regulations on key traits

To investigate a possible relevance of the observed omics variations for the genetics of key agronomic phenotypes or traits, we concentrate on the interplay between omics variables (*i.e.* synonymous and non-synonymous substitution rates, SNP density, gene expression and DNA methylation) for maize genes and their associated functions in biological processes (Figure 4). At the whole genome level in maize, we detected a weak negative correlation (r2=-0.27, pvalue<2.2e-16) between promoter methylation (in CG, CHG & CHH) and gene expression, whereas a weak positive correlation (r2=0.2, pvalue < 2.2e-16) has been recovered between gene expression and gene-body methylation in the CG context (Supplementary Figure 13a). However, when splitting omics variable into quantiles, clear omics interplay appears. Highly methylated genes (methylation > 50%) in the promotor are less expressed, while hypomethylated genes (first quantiles) correspond to the most expressed genes (Figure 4a). Gene-body methylation and expression level seem to be positively correlated (for the first quantiles) and negatively correlated for highly methylated genes (last quantiles, methylation > 58%). A negative correlation was observed between ka and gene-body methylation as well as expression with lower expression levels as the ka rate increases (Figure 4a), and with no clear pattern between ks or SNP density on gene expression (Supplementary Figure 13b). Interestingly, highly methylated and silenced genes belong to non-conserved (species-specific) genes whereas conserved genes do not show a clear association between gene expression and DNA methylation (Supplementary Figure 13c).

To study in more details the link between DNA methylation and gene expression, gene-to-gene analysis was conducted using the multivariate approach implemented in mixOmics in maize and *Brachypodium* (Rohart et al. 2017). Overall, samples were clustered by developmental stages and the transcriptomic (gene expression) and epigenetic variables (in mCG, mCHG and mCHH contexts), Supplementary Figure 14. Gene-to-gene analysis opened up possibilities to identify sets of genes with particular profiles, such as Hypomethylated/Upregulated (Hypo/Up) and Hypermethylated/Downregulated (Hyper/Down) genes, with expression expressed as TPM and methylation expressed as rpd, read per density (Figure 4b, Supplementary Figure 15). 441 and 50 genes were classified as Hypo/Up respectively in *Brachypodium* and maize and 440 and 604 genes classified as HyperDown respectively in *Brachypodium* and maize.

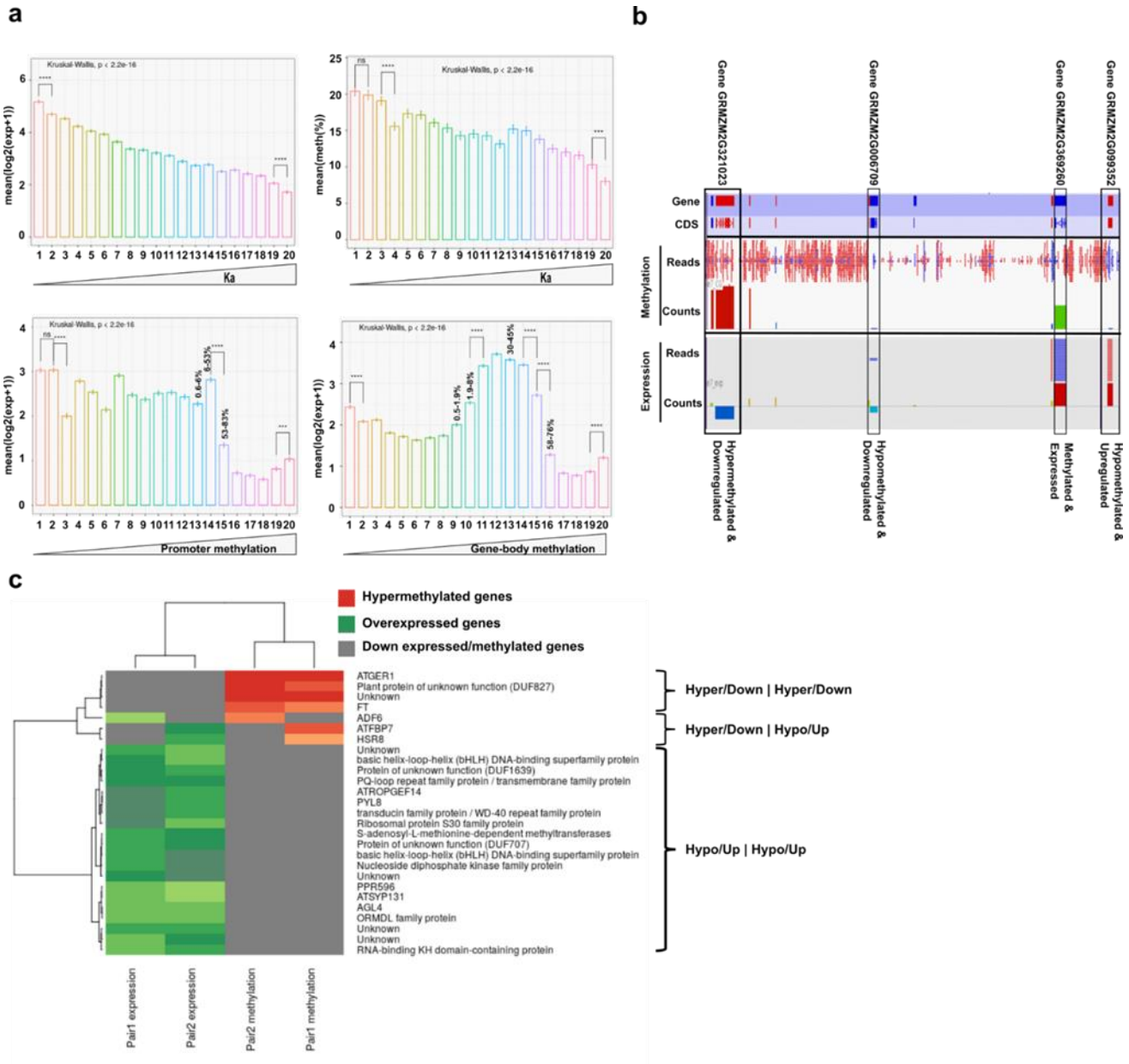


Figure 4: Multiomics regulation interplay in maize. a. Interplay between substitution rates (Ka expressed in quantiles at the top) and gene expression (top left) or gene-body methylation (top right). Bottom panel: link between DNA methylation level (expressed as in quantiles) and gene expression (expressed as log2(TPM)) for promoter (bottom left) and gene-body (bottom right). b. Examples of the link between gene expression and DNA methylation in maize for Hypo/Up (hypomethylated and upregulated) and Hyper/Down (hypermethylated and downregulated) genes. The coverage (top track, *i.e.* reads) and quantified (bottom track, *i.e.* counts) data are shown for methylation and expression of genes. Red for high expression / methylation levels and blue for weak expression / methylation levels. c. Methylation and expression profile of duplicated trait-driving genes in maize with contrasted signature in methylation and gene expression. Grey for not expressed / methylated genes; green for expressed genes and red for methylated genes. Hyper/Down indicates hypermethylated and down-regulated genes and Hypo/Up indicates hypomethylated and upregulated genes.

Using *Brachypodium* for GO analysis, Hypo/Up (441) genes (in CG, CHG & CHH) are enriched in GO terms like transporter, transmembrane, ligase, cation, ion, acid, etc., relating to signaling

processes while Hyper/Down (397) genes (in CG & CHG excluding CHH context due to the low number of genes impacted) are enriched in functions related to purine/ribonucleotide binding, hydrolase activity, ion binding, transporter activity, heat shock protein binding, etc. (Supplementary Figure 2d-e).

It has been proposed that polyploidization may promote the emergence of new phenotypes leading to species diversification, adaptation and domestication (Qi et al. 2021). To assess the impact of gene or block duplication divergence (in expression and methylation) on key traits, we focused on the recent maize WGD event. While expression divergence in most maize duplicates is not associated with methylation differences, 11% of the duplicated genes (out of 3,193) display a clear association between gene expression and methylation for at least one copy of the gene pair. In particular, 8% of the duplicated pairs had a hypomethylated and upregulated copy (referenced as Hypo/Up) while only 3% had a copy with the opposite pattern (hypermethylated and downregulated, i.e. referenced as Hyper/Down), Figure 4c, Supplementary Table 10. We then focused on pairs where both copies were contrasted in term of methylation and expression profiles (Figure 4c). Among them, 20 displayed Hypo/Up status, *i.e*. both copies of a pair are hypomethylated and upregulated. These included bHLH, ROPGEF14, SEP2/AGL4, ORMDL, PPR596, SYP131, PYL8, CCoAOMT1 (S-adenosyl-L-methionine-dependent methyltransferases) involved mainly in pollen/floral development, seed germination/growth and stress response. We found four gene pairs where both copies are highly methylated and not expressed (Hyper/Down), namely GER1/GLP1 (involved in fruit development), FT (flowering time) and two unknown genes. Finally, we found three gene pairs where the two copies show a contrasted omics regulation pattern, with one copy hypermethylated and downregulated, and the other copy hypomethylated and upregulated (Figure 4c). These genes are ADF6 (defense response), FBP7 (seed development) and HSR8 (cell wall & glucose response).

Discussion

Grasses as a key species complex to investigate polyploidization events

Polyploidization has been proposed as a driving force of plant genome evolution. However, the genomic reprogramming of duplicated compartments and genes inherited from polyploidization and underpinning such evolutionary success remains largely obscure. Polyploidization events are followed by a diploidization (or fractionation) process that consists of gene number reduction (Langham *et al.*, 2004; Thomas *et al.*, 2006; Woodhouse *et al.*, 2010; Schnable *et al.*, 2012; Murat *et al.*, 2014). The continuum of structural and regulatory modifications between the parental genome merger and the eventual return to a diploid status needs further investigation to decipher the biological functions gained from WGDs that may explain the role of recurrent polyploidizations during angiosperm evolution (Fox *et al.*, 2020).

To investigate the evolutionary trajectories of structural and regulatory patterns, we conducted a comparative omics analysis between eight modern grasses deriving from a 100 My old ancestor (AGK7) of 7 protochromosomes with 16K protogenes enriched in function related to transport activity. Despite widespread and recurrent polyploidization, most grass species have markedly few chromosomes. E.g., maize is expected to have n=28 chromosomes (7x2x2) based on polyploidizations since AGK7 alone, but in fact has only *n* = 10. The many examples of species with

far fewer chromosomes than predicted by past polyploidization events suggest a strong selective pressure that favors fewer chromosomes, although the advantages of a low chromosome number are not clear (Pont*et al.*, 2019; Escudero and Wendel, 2020). Extant grass genomes have derived from the inferred ancestor (AGK7 and AGK12) through distinct chromosomal fusions and fissions, as well as inversion events that tend to be located at the telomeres and more frequent in the polyploid context, without any reuse of synteny breakpoint, contrary to what has been proposed in animals (Maria Maggiolini *et al.*, 2020). Synteny breakpoints marking a recurrent inversion in chimpanzee and gorilla chromosomes homologous to the human chromosome 16 are near a conserved 23-kb inverted repeat composed of satellites, LINE and Alu elements. It is believed that this repeat mediated the inversion by bringing the chromosomal arms into close proximity, hence facilitating intrachromosomal recombination (Goidts *et al.*, 2005). Like in mammals, our analysis of the sequences surrounding inversions (on the wheat chromosome 6) supports the hypothesis of the involvement of sequence repeats in the emergence of inversions. For genes embedded in inversion, we observed higher non-synonymous substitution rate (in *Brachypodium* and maize), higher expression and lower methylation level (at promoter in CG-CHH-CHG contexts), suggesting that inversions play a role in genome plasticity during evolution. In the mimetic butterfly, a set of genes coadapted for mimicry was assembled by an inversion (Joron *et al.*, 2011). Dvorak *et al.* (2018) reported faster rates of genomic changes (inversions, translocations and duplications) in the *Triticeae* genomes compared to rice and sorghum, and faster substitution rates in the *Pooideae* branches compared to *Panicoideae* and *Oryzoideae*, with the highest rates in the A-, B- and *Ae*. *tauschii* genomes and the lowest rates in rice and sorghum. Such differences in rearrangement rates (particularly inversions) could be attributed to differences in lifespan, as suggested in mammals where the quickest rate of chromosome breaks was observed in mouse that has a short generation time (Murphy *et al.*, 2005). We propose here that such genomic rearrangements are also favored by polyploidization events, the merging of quasi-identical genomes in the nucleus, facilitating rapid differentiation of the two parental subgenomes necessary for correct homologous (and not homeologous) chromosome paring.

Devenir of duplicated genes and blocks following polyploidization

Deletion of duplicated genes does not occur at random but rather through a so-called subgenome dominance process, where most of the ancestral genes are preferentially retained on only one of the duplicated blocks. On the whole chromosome or genome level, this diploidization phenomenon then leads to a dominant subgenome enriched in singletons and characterized by gene retention (D or LF for 'least fractionated'), and a sensitive subgenome enriched in duplicates and characterized by preferential losses (S or MF for 'most fractionated') (Salse, 2012; Salse, 2016). Despite the general presence of subgenome dominance in paleo- or neo-polyploids, this phenomenon has not been observed in auto-polyploid plants, such as potato (Stupar *et al.*, 2007), poplar (Y., Liu *et al.*, 2017), and pear (Q., Li *et al.*, 2019).

In the current study, 25% and 23% of annotated genes were retained in pairs in the extant genomes, following the ancestral (ρ, 90 My ago) and maize-specific (5 My ago) WGDs, respectively. Pairs (or “diploidization-resistant” genes), as well as MF-located genes, are enriched in functional annotations such as transcription factor and transcription regulator, in contrast to singletons (and LF-located genes) enriched in signalization processes, in agreement with the findings of Freeling (Freeling, 2009). In *Brassicacae,* LF single-copy genes are reportedly enriched

in biological processes such as DNA repair and RNA interference, with the relevant genes preferentially deleted from MF1 and MF2 subgenomes (Yue Hao *et al.*, 2021). It was suggested that the affected mechanisms could subsequently reinforce the observed biased fractionation (Yue Hao *et al.*, 2021). The gene balance hypothesis tries to explain the differential retention of certain functional groups of genes in pairs by emphasizing the protein network level. Transcription factors and transcription regulators operate within protein-nucleic acid complexes at a particular dosage level, and the interaction with other proteins corresponding to duplicated genes (pairs or MF genes), makes them prone to being retained after polyploidization (Conant *et al.*, 2014; Freeling *et al.*, 2015)(Sémon and Wolfe 2007,. On the contrary, genes encoding proteins that function independently or are less associated with others in networks (proteins involved in DNA repair and metabolism, and RNA binding and interference), tend to retain only one of those copies and loose the redundant ones (Freeling, 2009). While the LF and MF boundaries are conserved at orthologous position between the investigated species, indicating that the genome fractionation has been established before the speciation, LF-MF fraction exchanges between pairs of ancestral chromosomes can trace homeologous chromosome exchanges during post-WGD evolution. Homeologous exchange is common during early polyploid formation, and large chromosome segments from one subgenome can replace another as observed in *Brassica* (Stein *et al.*, 2017), strawberry (Folta and Barbey, 2019), cotton (Page *et al.*, 2016), and millet (Shi *et al.*, 2019). In our present analysis of molecular evolution of duplicated genes and singletons, no clear distinction in non-synonymous substitutions (Ka) between pairs and singletons was determined, but pairs clearly displayed a significantly higher number of synonymous substitutions (Ks) compared to singletons. At the block level, LF-located genes showed significantly less synonymous substitutions than MF-located genes and non-synonymous substitutions ratios in favor of an excess of substitutions in MF-genes. Phylogenetic tree branch lengths confirmed this general pattern, where pairs have evolved faster than singletons in more recent MRCAs (< 27 My ago) and modern species. Overall, our results suggest that evolutionary speed is affected by both whether a gene belongs to the LF/MF fraction of the genome, and whether the gene is conserved in multiple copies (pairs) or not (singletons) after the WGD event. Pairs display a significantly higher number of substitutions and longer branch length compared to singletons, and LF-located genes evolve slower than the MF-located genes, jointly making the copies of gene pairs that are located in the LF compartments the most plastic fraction of the genome during grass evolution.

To what extent can this observed bias in molecular evolution between the LF and MF compartments be generalized across species? *Arabidopsis thaliana,* that experienced two polyploidization events (α and β, 24 et 40 My ago), retained between 23% and 29% of duplicated genes following the subgenome dominance process (Blanc *et al.*, 2003; Thomas *et al.*, 2006). The dominant subgenome (LF) has retained 70% of the genes since the paleoploidization event, whereas 46% and 36% of the genes have been retained in MF1 and MF2, respectively. In *Brassicacae,* an ancestral triplication (WGT) dating back to 15 My ago was followed by subgenome dominance and a reported loss of 60% of the ancestral triplicates, with 50%, 65% and 70% lost in the FL, MF1 and MF2 compartments, respectively (Cheng *et al.*, 2012; Liu *et al.*, 2014; Parkin *et al.*, 2014). Among the genes that derive from the WGT, 13.42% of polyploidy-derived genes accumulate more transposable elements and non-synonymous mutations than other genes during evolution. Although no biased fractionation between subgenomes was detected in the allotetraploid Ethiopian cereal teff, a general subgenome dominance in the expression atlas

across tissues was reported, with the B subgenome having overall higher homoeolog expression levels, perhaps due to its smaller size and fewer transposable elements (VanBuren *et al.*, 2020). In maize, subgenome dominance following the lineage-specific duplication dating back to 5 My ago, documented in the current analysis, is in line with previous results based on 28,1% of genes retained in pairs, reporting that LF blocks contain more expressed genes, less transposable elements and accumulate fewer substitutions (Schnable *et al.*, 2011; Renny-Byfield *et al.*, 2017; Zhao *et al.*, 2017). It has been reported that duplicated genes retained in rice are enriched in single nucleotide polymorphisms (SNPs) encoding less radical amino acid changes, suggesting that such "advantageous" material/genes inherited from WGDs are highly stable (i.e., slow rate of evolution) over the time (Chapman *et al.*, 2006).

Omics reprogramming following polyploidization

Duplicated genes and blocks have been proposed as a reservoir of alternative variants of gene regulation in terms of expression and methylation. In the current analysis, we established that half of the duplicated genes (after the ancient and recent WGDs) display gene expression and DNA methylation differences, and that pairs display less methylation, less expression, more SNPs (in maize for both ancient and recent polyploidization events), lower gene body methylation level (CG-CHH-CHG contexts) and higher rates of synonymous and non-synonymous substitutions (except for the recent duplication in Maize), compared to singletons. At the post-polyploidization block level, MF-genes show less expression with no methylation bias compared to LF genes. Such divergence in expression and methylation between duplicated blocks and genes has been proposed as a source of subfunctionalization (partitioning of ancestral functions between the duplicated genes) and neofunctionalization (gain of a new non-ancestral function in one duplicate), both being key forces of evolutionary plasticity in plants (Zou *et al.*, 2009). To what extent can such differences in regulation between duplicated blocks and genes be considered as a general post-polyploidization phenomenon?

In *Arabidopsis thaliana,* 74% of genes are conserved in pairs and 3% display difference in expression (Duarte *et al.*, 2006; Coate *et al.*, 2020) with singletons more expressed that pairs (Wang *et al.*, 2012). In cotton, LF-located genes are reportedly more expressed than the MF-located genes, and are associated with more TEs and siRNAs that may play a role in reducing or modulating the expression of flanking genes (Renny-Byfield *et al.*, 2015). The same authors investigated the expression of pairs in three tissues (petals, grains and leaves) and established that among 2000 gene pairs deriving from an ancestral duplication, more than 99% display expression difference between at least one of the tissues and 93% between the three tissues. Among 1,971 pairs active during the grain development (at 10, 20, 30 and 40 days post-anthesis, DPA), 84% display expression differences (Renny-Byfield *et al.*, 2014). In maize, 65% of pairs show expression differences on average across eight tissues (Schnable *et al.*, 2011), with expression dominance for 60% of the pairs in favor of one subgenome and 40% for the other, depending on the tissue considered (Zhao *et al.*, 2017). However, no expression difference between maize subgenomes (Li *et al.*, 2016), and no methylation difference (Renny-Byfield *et al.*, 2017) has also been reported. Nonetheless, the dominant (LF) subgenome has consistently been observed to have lower methylation upstream of genes, in particular lower CHH methylation around 500 bp upstream of the TSS (Renny-Byfield *et al.*, 2017; Zhao *et al.*, 2017). Overall, the LF fraction in maize

has been demonstrated to have more tandem gene duplications (Edger *et al.*, 2019), higher gene expression (Schnable *et al.*, 2011) and lower DNA methylation (Woodhouse *et al.*, 2014). In soybean, that experience a lineage-specific duplication 5–13 My ago, 80% of duplicated genes are conserved with 40% of them showing expression difference but no bias toward any of the subgenomes (Zhao *et al.*, 2017). The duplicated genes from the two subgenomes showed no differences in the distance to the nearest TE or in methylation levels (Zhao *et al.*, 2017). Although only a few genes in the soybean genome have a TE nearby (<1 kb), highly expressed soybean genes tend to be further away from TEs (Zhao *et al.*, 2017). In *Brassicaceae*, LF-located genes are more expressed, less methylated and accumulate less SNPs than MF-located genes (Cheng *et al.*, 2016). Parkin *et al.* (2014) reported that among the triplets inherited from the ancestral triplication in *Brassica*, 83% are differentially expressed in the leaf, with LF-located genes being more expressed but showing no difference in DNA methylation. *Brassica napus*, derived from a hybridization between *Brassica rapa* (An) and *Brassica olerace (Cn)* some 12,500 years ago, consists of MF (An) and LF (Cn) subgenomes, with 58% of pairs showing no expression difference in leaves and roots and no subgenome dominance in expression bias for the remaining pairs associated with expression differences (Chalhoub *et al.*, 2014). Li *et al.* (2020) investigated four tissues to conclude that between 70% to 85% of the retained pairs have a bias in expression toward the An subgenome depending on the tissue considered. Globally, CG, CHG and CHH methylation of genes and LTR retrotransposons are higher in the dominant BnC (LF) subgenome compared to the BnA subgenome, similarly to the results from natural *B. napus* (Chalhoub *et al.*, 2014). Zhang *et al.* (2021) confirmed the previous analysis in showing that singletons are less expressed and more methylated than pairs, and that singletons in Cn (LF) are more methylated than singleton in An (MF). In pear, that experienced a polyploidization some 30 My ago, duplicated blocks have been reported with no difference in gene loss, substitutions, expression and methylation (Q., Li *et al.*, 2019). Singletons in pear are more expressed with transcripts detected in a wider range of tissues, more methylated (CG in promoters and CHG in promoters and gene bodies), and have more substitutions (Ka) as well as Ka/Ks, compared to pairs. On the other hand, 54% of retained pairs show expression differences. In *Nelumbo nucifera*, that experienced a paleotetrapolyploidization 60 My ago, MF-located genes are more methylated and more expressed than LF-located genes, while singletons are more expressed in a wider range of tissues, more methylated (gene body), and have less substitutions compared to pairs (Shi *et al.*, 2020). In *Mimulus peregrinus* resynthesised and natural polyploids, the dominant subgenome had lower TE density and tended to have equal or lower expression compared to the nondominant subgenome (Edger *et al.*, 2017). Extensive homoeolog expression bias is also observed in hexaploid wheat (Ramírez-González *et al.*, 2018) and tetraploid *Tragopogon mirus* (Buggs *et al.*, 2010) and may be a common feature of recent polyploid grasses. In switchgrass (Lovell *et al.*, 2021), the K subgenome (LF) has higher gene density, more upregulated genes and lower substitution rates compared to the N subgenome, pointing to a stronger evolutionary constraint of the K subgenome and suggesting that the potential for adaptive evolution may be differentially partitioned between subgenomes. In the *Cucurbita* genus (Sun *et al.*, 2017), with an allotetraploidization that happened between 3-26 My ago, the two subgenomes have retained similar numbers of genes, and neither subgenome is globally dominant in gene expression. Integrating studies from 20 angiosperm species, De Smet *et al.* reported that singletons are more methylated and more expressed than pairs (DeSmet *et al.*, 2013). Wang *et al.* (2017) investigated

CG methylation of pairs in rice and concluded that genes showing methylation divergence also show expression differences. Investigations of the expressional dynamics of grass duplicates deriving form a 90 My ago paleotetraploidization event suggest that 57.4% (Yim *et al.*, 2009) and up to 85% (Throude *et al.*, 2009) of rice paleoduplicates have diverged in expression. In rice, retained ancient gene duplicates associated with high expression tend to have higher CG body methylation (Wang *et al.*, 2013), suggesting a direct role of epigenetic regulation in structural and expressional maintenance of duplicates, preventing pseudogenization, silencing and deletion, and ultimately retaining WGD-derived genes. It has also been reported that dominantly expressed genes in D/LF subgenomes have fewer 24-bp RNAs in their 1-kb flanking regions compared with their S/MF paralogs (Woodhouse et al. 2014).

Omics regulation interplay in grasses

In grasses, we provided evidence of the link between nucleotide substitutions, DNA methylation and genes expression. We report a negative correlation between Ka and both gene expression and DNA methylation, highlighting the interplay between (non-synonymous) nucleotide substitutions and gene regulation. In mollusk (Crassostrea giga) Song et al. (2018) reported a negative relationship between gene expression and non-synonymous sequence diversity, which suggests that purifying selection has played an important role in shaping genetic diversity. Highly expressed genes are often subject to strong selective constraints, and tend to evolve more slowly (Liao and Zhang, 2006, Park et al. 2012). While negative correlation between gene expression and non-synonymous substitutions is more obvious, the negative correlation between gene-body methylation and non-synonymous substitutions reported in the current study is more surprising. Most other studies found positive association of synonymous substitutions with gene-body methylation (Tsunoyama *et al.*, 2001; Glastad *et al.*, 2016; Lian *et al.*, 2020). The negative link between gene-body methylation and non-synonymous substitutions in grasses found here may suggest that DNA methylation plays a protector role by preventing non-synonymous mutations, despite the potential for increased mutation rates in methylated CpG dinucleotides (Chuang *et al.*, 2012; Monroe *et al.*, 2022). Indeed, methylated genes have been suggested to be more functionally important than unmethylated genes, and tend to evolve slowly (Takuno and Gaut, 2012; Sarda *et al.*, 2012). Another explanation could be that methylated genes that are functionally important such as the housekeeping genes, may be under purifying selection, and therefore have less nonsynonymous mutations.

To go further, we also examined the link between DNA methylation and gene expression. DNA methylation is a heritable epigenetic modification that has been shown to impact gene expression in plants and animals (Zhang *et al.*, 2008). In *Arabidopsis*, DNA methylation of coding regions is associated with genes that are expressed at medium to high levels and enriched for housekeeping functions, while methylation of promoter regions is generally associated with gene repression or silencing (Zhang *et al.*, 2006; Bewick and Schmitz, 2017, Ballinger et al. 2007). While no whole-genome correlation between DNA methylation and gene expression was detected in the current study, we show here for a subset of genes a negative link between promoter DNA methylation and gene expression and supporting the idea that DNA methylation play a key role in gene silencing but only for the highly methylated ones (Suzuki and Bird, 2008). We report a clear methylation dosage effect on gene expression with strong inverse correlation between promoter

methylation and gene expression, with major reduction in gene expression (up to silencing) when promoter methylation reaches above 50%. In respect to the DNA methylation of gene bodies, which has been reported to be positively correlated with gene expression, we demonstrate here that this relationship depends on the methylation level, with high gene-body methylation (>58%) associated with transcriptional repression. Finally, we highlighted that these interplays concerned only species-specific genes (and not conserved genes) which might suggest species specificity in gene expression regulation mediated by DNA methylation modifications.

Emergence of key traits through polyploidization

What are the consequences of polyploidization in biological and phenotypic innovation? Reports highlight the role of polyploidization on species diversification, crop domestication and in the establishment of important agronomic traits as well as the evolution of stress resistance (Hancock, 2005; Renny-Byfield and Wendel, 2014; Hannweg *et al.*, 2016; Hias *et al.*, 2018). In maize, the dominant subgenome contributes more to trait heritability than the nondominant subgenome (Renny-Byfield *et al.*, 2017) and in strawberry, the dominant subgenome largely controls several biological pathways related to agriculturally valuable traits like fruit flavour, colour, and aroma (Edger *et al.*, 2019). A recent study showed that subgenome-specific selection of defense response genes contributes to the environmental adaptation of allopolyploid *Brassica napus* (Lu *et al.*, 2019). In *Brassicaceae*, large-scale resequencing revealed that parallel selection of homoeologous genes derived from polyploidization is associated with morphotype diversification (leafy head) in *B. rapa* and *Brassica oleracea* (Cheng *et al.*, 2016). In *Brassicaceae*, GO terms related to the phosphate biology process and root development were enriched in over-retained genes following the ancestral WGT. In the cotton genome, a search for genes responsible for long white fibers revealed 620 homoeologous pairs that have been subjected to domestication selection either in the A or D subgenome, while only 34 homoeologous pairs exhibit selection signals in both subgenomes, indicating that the coexisting subgenomes have been under asymmetrical domestication selection (M., Wang *et al.*, 2017). By comparing the distributions of positively selected genes in cotton as well as fiber-related quantitative trait loci, Zhang et al. (2015) concluded that the A subgenome was selected for fiber improvement genes, and the D subgenome was selected for stress tolerance genes. Polyploidy has been also proposed to give rise to new aromatic profiles of strawberry fruits (Ulrich and Olbricht, 2013). In lupin, a dinitrogen (N_2)-fixing legume that evolved from a whole-genome triplication (WGT) event, the number of phosphorus-use efficiency (PUE) genes has increased through WGT, tandem, and dispersed duplications, with WGT-derived genes enriched in GO terms related to the phosphate biology process and root development (Xu *et al.*, 2020). In switchgrass, 75.9% of biomass SNP–heritability was attributable to the N subgenome, and only 24.1% to the K subgenome across 10 common garden experiments (Lovell *et al.*, 2021). In our study, we were able to identify key genes of agronomic importance with particular omics patterns when assessing methylation and expression signature of specific and recently duplicated genes in maize. Most of the duplicated genes with an identical methylation and expression signature for both copies were hypomethylated and expressed, and involved in seed development and growth. This is the case for: i) ORMDL family gene, where reduction in gene expression resulted in sterile phenotype with abnormal pollen morphology and staining in rice (Chueasiri *et al.*, 2014); ii) AGL4/SEP2, where reduction in SEP activity led to the loss of normal ovule development (Favaro *et al.*, 2003) and can subsequently

influence the architecture of the inflorescence (Ma *et al.*, 2022); iii) SYP131, where triple mutant syp124 syp125 syp131 resulted in gametophytic defect (Slane *et al.*, 2017); iv) bHLH, a transcriptor factor involved in plant growth and development (X., Zhou *et al.*, 2020; Yaqi Hao *et al.*, 2021); v) PYL8, where overexpressing plants exhibit hypersensitive phenotype to ABA in seed germination, seedling growth and establishment (Lim *et al.*, 2013); vi) ROPGEF14, where ropgef1, ropgef9, ropgef12 and ropgef14 quadruple mutant showed reduced pollen tube elongation (Chang *et al.*, 2013); etc. We show that all these genes are duplicated in maize and expressed during grain development with no or low methylation level for both pairs. For three maize gene pairs, one copy was hypomethylated and overexpressed, while the other copy of the pair was hypermethylated and underexpresses. This was the case of: i) ADF6 where down-regulation in cotton rendered the plant tolerant to *V. dahlia* infection (Sun *et al.*, 2021); ii) FBP7, where down-regulation resulted in maternally controlled defects in seed development in Petunia (Colombo *et al.*, 1997; Cheng *et al.*, 2000); iii) HSR8, involved in sugar responsive growth and development (Li et al. 2007) or stress tolerance (Zhao *et al.*, 2019). Finally, four maize gene pairs were hypermethylated and silenced for both copies. Two of them are unknown genes and the remaining two correspond to FT and GER1/GLP1 genes. FT modulates flowering transition and inflorescence architecture (Wickland and Hanzawa, 2015) while GER1 is involved in stress response (Ilyas *et al.*, 2016) and is expressed during fruit development and ripening in plum (El-Sharkawy *et al.*, 2010). These observations open up new possibilities for exploitation of DNA methylation in breeding, where manipulation of the methylation pattern of duplicated genes (e.g. by gene editing techniques) could bring about changes in gene expression that determine agronomically important traits, such as yield and resistance to biotic and abiotic stresses.

Model of polyploidization-driven genomic plasticity

In the figure 5a, we propose a model of possible association between WGD and the potential consequences on genome reprogramming based on (i) significant differences observed between duplicated regions or genes, and (ii) consistent differences observed across species and WGD events (figure 5a). Following such rules, our model of polyploidization consequences concludes that (1) inverted genes have higher substitution rates, are less methylated and more expressed [referenced as #1 in Figure 5b], (2) conserved genes have lower substitution rates, are less methylated and more expressed [referenced as #2 in Figure 5b], (3) pairs have more substitutions, are less methylated and less expressed [referenced as #3 in Figure 5b], (4) LF-located genes have less substitutions with no clear consensus on expression or methylation bias between LF/MF-located genes [referenced as #4 in Figure 5b], (5) more than 50% of the pairs show within-pair expression differences [referenced as #5 in Figure 5b], MF-located paired genes enriched in transcription factor and regulator activity [referenced as #6 in Figure 5b].

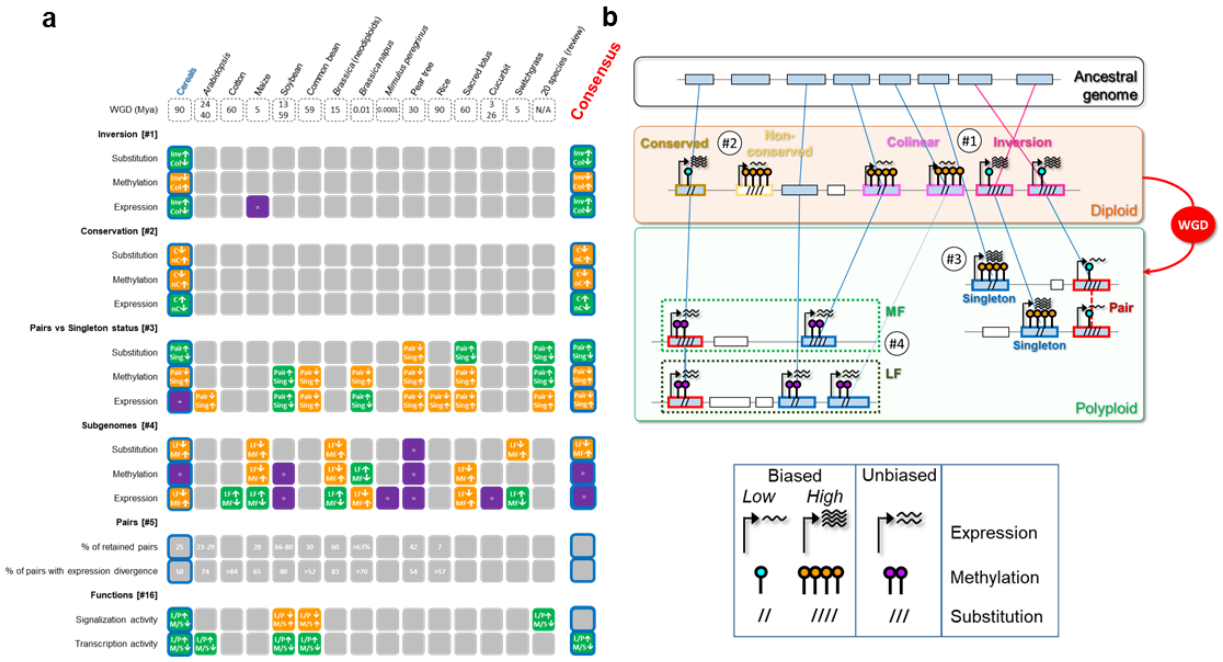


Figure 5: Model of post-polyploidization genomic reprograming. a. Comparative analysis of the omics variation reported for inverted, conserved, duplicated, LF-/MF-located genes (lines) from the current analysis (first column) and the literature from 51 angiosperm species. When similar conclusions are made in other species compared to the current study, the color of the first column is applied to the corresponding columns. When opposite conclusions are made, different color is used. No observed differences are indicated in purple. Consensus conclusions are drawn for omics trends predominantly observed in angiosperm paleo- and neo-polyploids. b. Schematic model of post-polyploidization genome reprogramming for inverted, conserved, duplicated, LF-/MF-located genes, with genes shown as colored rectangles and substitutions, expression and methylation according to the legend at the top right.

Structural and functional postpolyploidy changes driven by subgenomic dominance seem to require some time to "evolve" and "stabilize", as suggested by the current analysis as well as by results on resynthesized polyploids (Renny-Byfield *et al.*, 2015). Consequently, such genomic reprogramming following polyploidization may lead to novelty at (1) the functional level (neo- and subfunctionalization, Lynch and Conery, 2000; Wang *et al.*, 2012, (2) the network level with the maintenance of a stoichiometric balance of gene product interactions (or connectivity) in macromolecular complexes (Birchler and Veitia, 2011; Birchler and Veitia, 2014), (3) the phenotypic level with an heterosis effect with transgressive performances (Birchler *et al.*, 2010), (4) the allelic level in masking deleterious recessive mutations (Gu *et al.*, 2003), (5) the adaptation level with escape from adaptative conflicts (Des Marais and Rausher, 2008), (6) the regulation level with novel expression and methylation patterns (Seoighe and Wolfe, 1999; Aury *et al.*, 2006; Yang and Gaut, 2011), all potentially contributing to a new polyploid machinery absent from the diploid progenitors. Looking from the perspective of evolutionary outcomes backwards, it appears that postpolyploidization changes are established to strike a balance between stability and novelty. On the one hand, the structural and functional partitioning of the subgenomes reinforce the differentiation of homoeologous chromosomes (subgenomes), leading to stabilization of meiotic pairing and increased fertility of the nascent polyploids by prevention of

homoeologous pairing. On the other hand, these changes provide genetic redundancy 'geared' toward phenotypic plasticity and novelty that could by advantageous in many ecological contexts. However, we need to be cognisant of the fact that by looking at the extant polyploids, we are only assessing the successful polyploid lineages and ignoring an unknown number of polyploidizations that disappeared as evolutionary dead ends, potentially creating an illusion that polyploidy is always associated with evolutionary advantage. We can speculate that striking the right balance between stability and novelty after WGD is rare, and the extant paleopolyploid lineages achieved their evolutionary success through independent trajectories, albeit showing some level of convergence. This view is consistent with the results presented here, demonstrating that diverged polyploid lineages sharing a common WGD event often present the same patterns of structural changes and evolutionary dynamics, but these patterns are difficult to generalize across independent WGD events. The lack of general patterns could be partially due to confounding non-WGD factors (differences in TE and recombination landscapes, epigenetic regulation, GC-content, efficiency of DNA repair mechanisms, founder effects, selection constraints and other population-genetic forces) operating over millions of years after polyploidization. Nonetheless, a common mechanism of biased fractionation remains enigmatic, and it appears that polyploids harness the potential of genome duplication, at least partially, in lineage-specific ways. Given that probably all grasses are derived from the ρ paleopolyploidization, it is impossible to imagine the diversification and the very existence of this clade without some relation to WGD. In this sense, WGD is unequivocally linked to the evolutionary success of grasses during the past 100 My; although it remains difficult to attribute this success to particular genomic consequences of polyploidization. Overall, the current study clearly demonstrates that post-polyploidization reprogramming is more complex than the traditionally reported differentiation of subgenomes and singletons-pairs. While statistically significant differences in gene expression, DNA methylation and substitution rates can be identified in various post-WGD fractions of the extant genomes, patterns that are strictly consistent across different clades and WGDs are rare, with the clearest distinction in LF-MF substitution rates. The lack of generalized patterns of WGD consequences highlights our poor understanding of polyploidy-driven evolution and calls for a critical and comprehensive comparison across independently polyploidized lineages. Nonetheless, we demonstrate that a detailed characterization of omics profiles across duplicated gene copies can prove useful for our understanding of trait evolution in the polyploid context, particularly in polyploid crops where it could identify new targets for breeding and improvement.

Funding

Completion of this article was supported by the ‘Région Auvergne-Rhône-Alpes’ and FEDER 'Fonds Européen de Développement Régional' (#23000816 project SRESRI 2015), the Institut Carnot Plant2Pro (#0001455 project SyntenyViewer 2017), The ISITE CAP2025 (#00002146 SRESRI 2015 ‘Pack Ambition Recherche Project’ TransBlé 2018).

Materials and Methods

karyotype reconstruction, evolution and rearrangements.

Genomes - The different genomes used for this study are Brachypodium distachyon (phytozome v9), Oryza sativa (phytozome v11), Sorghum bicolor (phytozome v9), Zea mays (phytozome v11), Setaria italica (phytozome v8), Triticum aestivum (Chinese spring v1.0), Triticum dicoccum (Zavitan WEW v1.0), Hordeum vulgare (Morex v1.0). **Ancestral karyotype reconstruction -** AGKs were reconstructed according to a two-stage procedure. While the ancestral karyotype is reconstructed in the first stage, the ancestral gene content of such karyotypes, and the gene order, are inferred in the second stage. This two-steps approach is crucial to obtain accurate ancestral genomes that are not fragmented into a large number of conserved syntenic blocks (or CARs). Stage 1, in which the ancestral karyotype (*i.e.* the ancestral chromosome number) is determined, is based on genome alignments (using CIP or cumulative identity percentage and CALP or cumulative alignment length percentage blast parameters (Salse et al. 2009). Conserved genes (*i.e.* putative protogenes or pPG) are identified from these alignments and pPGs conserved in all investigated species (*i.e.* core protogenes or core-pPG) are then extracted. These core-pPGs are used to identify synteny blocks (SBs) using DRIMM synteny software (Pham and Pevzner, 2010), removing groups of fewer than five genes. SBs are then merged using MGRA (Alekseyev and Pevzner, 2009) software based on chromosome-to-chromosome orthologous relationships between the compared genomes. Stage 1 thus yields the ancestral protochromosomes (also referred to as CARs), corresponding to independent sets of blocks sharing paralogous and/or orthologous relationships in modern species. Stage 2 of the procedure is aiming at ordering protogenes on the previously defined protochromosomes. Genes (pPGs) conserved between pairs of species but not constituting core-pPGs (used in stage 1) are integrated within SBs with DRIMM synteny software and then mapped onto the protochromosomes, delivering an exhaustive set of ordered protogenes (oPGs). CARs then correspond exclusively to diagonals in dotplot-based comparative genomics deconvolutions of the synteny between the investigated species, used here as a validation procedure. While in step 1, dotplot diagonals are identified based on core-pPGs, in step 2, such diagonals are enriched using pPGs. Putative orthologous (or ancestral) genes that have been transposed outside CARs (no longer part of dotplot diagonals) in the course of evolution are not identified and considered for the ancestral genome reconstruction. Our approach also allowed the reconstruction of a pre-WGD AGK ancestor by merging the post-duplication regions into a pre-duplication CAR. In this particular case, following the two stages approach described previously, the karyotyping stage (1) is performed using genes retained as pairs in the post-WGD ancestors, and the enrichment stage (2) is performed using the remaining singletons, as duplicated genes may have returned to singletons due to fractionation. In this particular case, singletons from both paralogous blocks are intercalated between conserved paralogs according to their current positions in the modern genomic regions (if no outgroup is available) or according to their conserved order within synteny blocks if outgroups are available. Following this strategy, the pre-WGD AGK was then constructed on the basis of paralogs (*i.e.* pairs in post-WGD ancestor defining a unique position in the pre-WGD ancestor) with CIP/CALP, DRIMM-synteny, MGRA tools described above, followed by an enrichment of the pre-WGD protochromosomes with singletons intercalated between conserved paralogs (*i.e.* each singleton located within a physical interval in the pre-WGD ancestor flanked by retained pairs).

Synteny breakpoints characterization - Alignment of the protogenes from the post-ρ ancestor (AGK12) to the genes of extant species allowed building an atlas of synteny breakpoints (i.e. ancestral chromosome fusion-fission sites). A transition between two ancestral chromosomes on an extant chromosome defined an SBP region bounded by two extant genes, orthologs of protogenes from two distinct ancestral chromosomes. Despite ancestral chromosome fusions-fissions, we investigated inversions detected from alignments of ordered protogenes of the post-ρ AGK12 and the gene order in the extant species investigated. Inversions were detected by analyzing the relative positions of genes on chromosomes of extant species compared to the inferred AGK order. Each gene has been tagged as non-inverted (collinear) or inverted relative to AGK. Inherently an inversion consists of the inversion of the order of two consecutive genes in respect to the ancestral gene order. In order to focus on large inversions, we set-up a minimum threshold of 10 inverted extant genes compared to AGK12 to define an inversion. Inversions were clustered into inversion families based on the value of their Jaccard index that gauges the similarity in terms of AGK gene content between inversions. Two inversions were included in a family when their Jaccard index value was 0.6 or above. Relying on inversion families, inversions have been refined by merging the inversions initially detected when belonging to two families included to a third one. A family was considered as included in another family when 90% of its AGK gene content is present in the second one. After merging of the inversions, a new step of inversion family clustering has been performed. In the end, based on the composition of the families, we have clustered the inversions according to their evolutionary origins and placed them on the phylogenic tree of grasses. For each branch of the phylogenic tree, the rate of inversions was calculated as the number of inversions over evolutionary time in million years. For each investigated species, we performed 1000 simulations where we randomized the position of the observed inversions along the genome (thus keeping the species specific inversion number and size distribution constant) and subsequently computed the gene density (in number of genes per Mb) for the randomized inversions. **Genome fraction inference -** In each extant genome, dominance at the gene level was inferred by comparing counts of descendant genes in different extant chromosomes over non-overlapping windows of 100 ancestral genes from the pre-WGD ancestral grass karyotypes (AGK7) for the 8 investigated species and AGK12 for maize and wheat. For each extant genome, we categorized the genes as LF, MF and non-significant. For each ancestral window of 100 genes, the number of ancestral genes retained in the two corresponding duplicated blocks in the extant species where compared. Significantly biased retention corresponded to a p-value < 0.05, where the p-value is the probability of observing at least this difference between two windows based on binomial distribution (where n is the windows size and p is the mean of the compared counts divided by n). This procedure was repeated on the genome of Maize, tetraploid wheat and hexaploid wheat based on AGK12 gene order to detect the LF/MF genes in these species.

Cereal phylogenetic history reconstruction.

Homology detection and sequence alignment - We detected homology between the cereal genes by performing an all versus all blast (Altschul *et al.*, 1990) whose output was processed using Silix (Miele *et al.*, 2011), followed by Hifix (Miele *et al.*, 2012).The amino-acid sequences of the generated homologous families were aligned using muscle (Edgar, 2004). Gblocks (Castresana,

2000) with options -b5=h -b4=2 was then used to extract conserved blocks from the multiple sequence alignments. This step identified 75 families presenting either no conserved blocks or at least one sequence absent from the detected conserved blocks. These cases likely resulted from spurious homology assignment and we decided to further refine them using a walktrap algorithm (Pons and Latapy, 2006), where the inverse of the pairwise poisson distance was used to give weights to the edges between proteins. This step yielded 175 additional homologous families with conserved blocks. The complete procedure resulted in 22,129 aligned homologous gene families (containing a total of 317,660 genes from the investigated species). **Reconciled gene tree inference -** Given the limited divergence of the species under investigation here (*e.g.*, potentially <10 000 years have passed since the divergence between the A subgenomes of T.*dicoccum* and T.*aestivum*), we chose to perform gene tree inference by jointly taking into account sequence and reconciliation information, which have been demonstrated to result in better quality gene trees (Szöllosi *et al.*, 2013; Scornavacca *et al.*, 2015). Our specific approach here most closely resembles that of TERA (Scornavacca *et al.*, 2015), but with the added possibility of a non-binary species tree (which is needed in our case as the separation between the A, B and D wheat lineage is unresolved and/or complicated by reticulation events (Glémin *et al.*, 2019) and whole genome duplications (two recent WGDs in the case of hexaploid wheat). For each gene, a distribution of 1000 bootstrap trees were inferred using iqtree (Nguyen *et al.*, 2014; Hoang *et al.*, 2018). This distribution was then used in our reconciliation procedure to compute the reconciled gene tree that minimizes the joint reconciliation-sequence score. After this, optimal branch lengths were computed for the inferred topologies using gene families' codon alignment and GTR-GAMMA from RAxML (Stamatakis, 2014). **Computing Ka and Ks -** Codon multiple sequence alignments were computed from the protein alignments using pal2nal (Suyama *et al.*, 2006). The number of substitutions per synonymous site and per non-synonymous site (Ka and Ks, respectively) were computed using the kaks function of the seqinr package (Charif and Lobry, 2007). The Ka or Ks profiles of different sets of genes were compared using the medians of the main peak of the distributions, which we evaluated as a more reliable measure than the distribution modes. The statistical significance of these differences was tested using a bootstrap procedure. Unless indicated otherwise, 1000 bootstraps were performed for each comparison. **Inference of substitution rates in Triticeae -** In order to gain further insight into the evolution of wheats, we directed particular attention on *Triticeae* species. We extracted 3 905 gene families in which *Triticeae* (namely here Barley, Wheat4xA, Wheat4xB, Wheat6xA, Wheat6xB, Wheat6xD, T.*urartu* and A.*tauschii*) presented exactly one ortholog each. Recently (Naser-Khdour et al., 2019) showed that model violation in phylogenetic reconstruction was both prevalent and deleterious among partition-based phylogenetic analysis. Indeed, most models of sequence evolution make the assumption that sequence evolution is stationary, reversible and homogeneous (SRH conditions). Accordingly, we tested each of the 3,905 gene family DNA alignments for violating the SRH assumptions, using scripts adapted from (Naser-Khdour *et al.*, 2019) . 930 gene families exhibited signs of SRH conditions violation and were thus excluded from our study. The remaining 2,975 gene family DNA sequences were used to construct a partitioned multiple alignment, upon which the best tree topology and branch lengths were inferred using RaxML (Stamatakis, 2014), with 100 bootstraps (the topology was constrained so that only the speciations of the wheat A, B and D lineages were left uncertain). The optimal model for each partition was inferred using iqtree (Nguyen et al. 2015). **Comparative SNPs analysis**- For maize, unimputed diversity data within the

third generation Zea mays haplotype map (Bukowski *et al.*, 2018) were downloaded from public sources. Individuals with outlier values [>3rd quartile+(1.5* interquartile range)] of missingness were removed, together with sites showing outlier values of total depth. For Brachypodium, a custom vcf file was created as follows. Chromosome-level assemblies for 54 lines of B. dystachyon (Gordon *et al.*, 2017) were downloaded from public sources and shredded into 300bp fragments with 50bp sliding step using the GenomeTools shredder command. The shredded fragments were converted into de-duplicated pseudo-fastq fragments with tally (Davis *et al.*, 2023) and mapped onto the common reference used throughout this study using bwa mem (Li, 2013). After coordinate-sorting (picard-tools SortSam) and addition of read groups (gatk AddOrReplaceReadGroups), vcf files were generated separately for each chromosome using GATK4.1.8 HaplotypeCaller, restricting the SNP calling to exon space only. For hexaploid wheat, unimputed SNP data produced by the WHEALBI initiative (Pont, Leroy, *et al.*, 2019) were used. For each species, per-gene SNP densities were calculated as the number of SNPs (>0.05 minor allele frequency) per 100bp of coding sequence, using the vcf files described above, relevant gene annotation files, and the commands intersect and groupby from the bedtools suite.
Expression data.

RNA-samples preparation - Seeds from *Brachypodium* (genotype BD21-3), maize (genotype B73) and wheat (genotype Recital) were grown and pollinated under standard conditions. Grains were sampled during the grain development at three stages (stage 1 for the cell division phase, stage 2 for the storage protein accumulation phase and stage 3 for the dehydration phase): 9, 16 and 28 DAP (days after pollination) in *Brachypodium*, at 7, 15 and 35 DAP in maize and at 100, 250 and 500°D (degree days) in wheat. These developmental stages have been determined to be equivalent by morphological, biochemical and genetic criteria. Each sample was a pool of at least 10 whole seeds collected from at least two different plants for RNA extraction. Grains (~1 g of tissue) were ground in liquid nitrogen and were extracted with 4.5 ml of buffer (10 mM Tris-HCl, pH7.4, 1 mM EDTA, 0.1 M NaCl, 1% Sodium Dodecyl Sulfate) and 3 ml of phenol – chloroform – isoamyl alcohol mixture 25:24:1. The supernatant was extracted one more time with the same phenol solution in order to eliminate proteins and starch. The nucleic acids were precipitated by addition of 0.1 vol of 3M sodium acetate pH 5.2 and 2 vol of 100% ethanol. After precipitation, RNA was rinsed one time with 70% ethanol and the pellets dissolved in RNase-free water. Purification was performed with a DNAse treatment RNase-Free DNase Set (Qiagen), followed by the RNeasy MinElute Cleanup Kit (Qiagen). The integrity of RNA was checked with an Agilent 2100 Bioanalyser microfluidics-based platform, using RNA 6000 Nano Chip kit and reagents (Agilent Technologies). **Control genes used for comparative grain kinetics -** Q-RT-PCR expression profiles have been produced with control genes known to be specifically expressed during cell division (*SUBTILISIN* gene corresponding to TaAffx.79990.1.S1 in wheat, GRMZM2G039538 in maize and Bradi1g08670-08450 in *Brachypodium*), filling (*Opaque2/SPA* gene corresponding to TaAffx.15974.1.S1 in wheat, GRMZM2G015534 in maize and Bradi1g55450 in *Brachypodium*) and dehydration (*Rab17* gene corresponding to TaAffx.58091.1.S1 in wheat, GRMZM2G079440 in maize and Bradi4g22280-22290-Bradi3g43870 in *Brachypodium*) phases of the grain development in grasses. RNA extractions were performed according to the ZR Plant RNA MiniPrep™ (Zymoresearch) protocol; DNase set (Qiagen) was used for RNA purification. When necessary, the residual DNA in the RNA samples was removed using the DNase set (Qiagen) and

RNeasy Minelute Cleanup (Qiagen) kits. cDNA synthesis was performed with a Transcriptor first strand cDNA synthesis kit (Roche) following the procedure described by the manufacturer with 0.5 µg of RNA in 20 µl of reaction. The thermal cycling conditions were 10' at 25°C/ 30' at 55°C/ 5' at 85°C. The reverse transcription reaction was diluted to a final volume of 200 µl, and 4 µl of synthesized cDNA was used as a template for real-time PCR using the LightCycler® 480 (Roche Diagnostics). The reactions were performed in 10 µl containing 1 x LightCycler® 480 DNA SYBR Green I Master (Roche Diagnostics). The quantification cycles (Cq) were analyzed using the LightCycler®480 software version 1.5.0 and normalized with reference genes using an 'Advanced Relative Quantification' profile to obtain a normalized ratio E_t^{-CqR}/E_r^{-CqR} (with CqT/CqR: cycle number at target/reference detection threshold (crossing point) and E_t/E_r: efficiency of target/reference amplification (10^{-1}/slope)). The specificity of the amplification was confirmed *via* melting curve analysis of the final PCR products by increasing the temperature from 65°C to 95°C. The PCR efficiency was calculated for each gene using a standard curve of serial dilutions and was used in the relative expression analysis. All observations were expressed as means ± S_D. **RNA library construction and sequencing -** The total RNAs were analysed with a capillary Shimadzu MultiNA microchip electrophoresis system. The total RNAs were first sheared with ultrasound (3 pulses of 30 sec at 4°C). From the sheared total RNA samples, the 3' polyA+ fragments were purified by means of oligo(dT) chromatography. An RNA adapter was then ligated to the 5´-phosphate of the 3' fragments. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and M-MLV H- reverse transcriptase. The resulting cDNAs were PCR-amplified using a high fidelity DNA polymerase. For Illumina sequencing, the cDNA fractions in the size range of 250 – 400 bp were eluted from preparative agarose gels. Aliquots of the size fractionated cDNAs were analyzed by capillary electrophoresis. RNA-seq was performed separately for the considered samples/libraries on illumina Hiseq (single reads of 100 bp). **RNA-sequences analysis -** RNA-seq reads were cleaned using CutAdapt (Martin, 2011) prior to mapping on the reference genomes using HISAT2 (Kim *et al.*, 2019) software with default parameters. Expressed genes were identified using TPM (Li and Dewey, 2011) threshold (>1). Differential expression analysis was done using edgeR package (Robinson *et al.*, 2010). Genes with less than five counts per million were dropped before library normalization using the trimmed mean of M-values (TMM) method. Count distribution was assessed with generalized linear model and dispersion estimated according to McCarthy et al. (2012). A gene was considered as differentially expressed (likelihood ratio test) if its adjusted p-value (Benjamini-Hochberg) was below 0.05. Genes were scored according to their expression profiles with, e.g., a profile 0-1-0 corresponding to an expression observed only at stage 2.

Methylation data.

WGBS library construction and sequencing - Whole-genome bisulfite sequencing was performed by the Integragen laboratory (Evry, France). Extracted genomic DNA was fragmented to approximately 200 base pairs (bp), and methylated adapters compatible with sequencing on an Illumina HiSeq instrument were ligated. The resulting libraries were then bisulfite converted (EpiTect Qiagen) and purified. Real time-PCR assay was used to determine the optimal number of PCR amplification cycles required to obtain a high diversity library with minimal duplicated reads. The sequencing was realized with paired ends (2×100bp) on an Illumina HiSeq™4000 platform with a minimal theoretical coverage of 15X. **DNA methylation analysis -** The quality of raw reads

was first checked with FASTQC (v0.11.7) before adapter trimming (illumina adapters), base quality (>28) and 5' trimming (6 bp, to reduce methylation calling bias) with Trim Galore (v0.5.0). Trimmed reads were then mapped to corresponding reference genomes, i.e. *Brachypodium distachyon* (Phytozome v9) and *Zea mays* (Phytozome v11) using Bismark (v0.22.3, Krueger & Andrews, 2011) under bowtie2 (v2.3.4.3, Langmead et al. 2009). All DNA methylation contexts (CG, CHG & CHH) were extracted using default options of the bismark workflow. The methylKit R package (v1.18.0) was used to annotate methylation data into gene features, *i.e.* promoter (500 bp around the TSS) and gene-body (exons + introns) for each methylation contexts. With methylation expressed in percentage, a new normalization approach for DNA methylation was used to assess the link between gene expression and DNA methylation. We call this normalization rpd, for reads by density: rbd = mCs x ratio (0-1) where ratio = mCs / (mCs + Cs). Differentially methylated genes (DMGs) were assessed using edgeR package with rpd data (Chen *et al.*, 2018; Robinson *et al.*, 2010). Methylation differences were assessed using the likelihood ratio test and p-values adjusted by the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR). Methylation differences were considered when adjusted p-values were below 0.05. Genes were scored according to their methylation profiles with, e.g., a profile 0-1-0 corresponding to methylation observed only at stage 2.

Statistical analysis - Statistical analyses were done with R software under Rstudio (R Core Team, 2018). Differences between evolutionary structural features were assessed with Kruskal-Wallis (or Wilcoxon), t-test and/or anova test. Tukey's post-hoc was used when significant results were computed with anova. Statistical tests were considered significant at p-value (FDR) < 0.05.

4 Résultats et Discussion

4.1 Evolution des génomes des graminées

4.1.1 Reconstruction des génomes ancestraux et du scénario évolutif des graminées

L'ensemble des travaux menés pendant cette thèse se basent sur la reconstruction des caryotypes ancestraux des céréales, *ancestral grasses karyotypes* (AGK), pré- et post-WGD ancestrale ρ. Les génomes ancestraux correspondent au contenu et à l'ordre des gènes conservés au sein des 8 espèces étudiées. Les AGK ont été inférés sur la base de l'analyse comparative du sorgho, de *Brachypodium* et du riz, ces trois espèces représentant respectivement les trois 3 sous-familles du panel, *Panicoideae*, *Pooideae* et *Ehrhartoideae*. Elles ont été choisies car elles n'ont pas subi d'évènement de polyploïdisation additionnels depuis la WGD ancestrale ρ. Les caryotypes ancestraux ont été reconstruits par Cécile Huneau en se basant sur la méthode décrite page 45 dont j'ai exploité les résultats pour l'analyse des réarrangements chromosomiques décrite dans la suite du manuscrit. Les résultats obtenus confirment et affinent les scénarios publiés par Murat *et al.* en 2017, et Pont *et al.* en 2019, qui identifient deux versions de l'ancêtre commun des céréales, pré- et post-WGD ancestrale. Le génome pré-ρ, noté AGK7, compte 7 protochromosomes. Ce nombre de chromosomes est concordant avec les précédentes reconstructions citées, mais le nombre de protogènes identifiés et ordonnés est de 10 286 ce qui représente un enrichissement par rapport aux 7010 et 8581 gènes précédemment répertoriés. Le génome post-ρ, noté AGK12, compte 12 protochromosomes et comporte 16 560 protogènes ordonnés, ce qui représente également un gain, quoique moins important, par rapport aux 14 241 et 16 464 gènes ordonnés dans les deux études citées. L'analyse des fonctions de ces gènes, ou *gene ontology* (GO), montre un enrichissement en gènes codant des activités transférases correspondant aux enzymes catalysant des transferts de groupement fonctionnel, méthyle ou phosphate par exemple, des activités de transports intra- et intercellulaire, et des activités de signalisation correspondant aux mécanismes de transfert d'information au sein de l'organisme. Ces fonctions sont à la base du métabolisme cellulaire de l'ensemble des plantes à fleurs, ce qui est attendu lorsque l'on construit un consensus des gènes ancestraux, qui par définition correspond à la fraction des gènes conservée chez la plupart des espèces modernes étudiées.

Disposant des génomes ancestraux AGK7 et AGK12, il devient possible de proposer un scénario expliquant les trajectoires évolutives des différentes branches de l'arbre phylogénétique des graminées depuis 100 millions d'années. L'évolution des caryotypes modernes par rapport aux génomes ancestraux peut être visualisée sous forme d'un dotplot illustrant la conservation des protogènes avec les gènes des espèces modernes. Ces alignements sont représentés par des dotplot projetant le code couleur des protochromosomes AGK 1 à AGK 7 sur les gènes modernes au sein de chaque espèce (Figure 30). Ces comparaisons entre génome ancestral et génomes modernes permettent également d'identifier les points de ruptures de la synténie (ruptures des diagonales), appelées '*synteny break point*' (ou SBP).

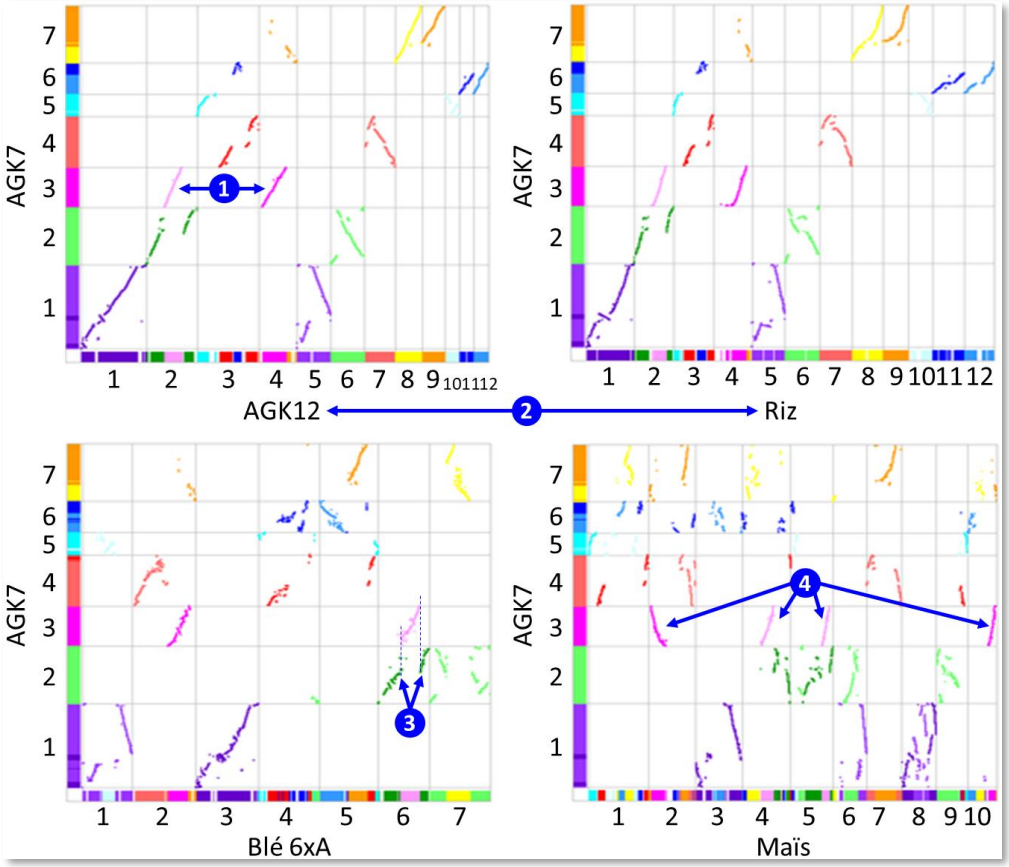


Figure 30 : dot plot représentants les relations d'orthologies entre AGK7 et AGK12, le riz, le sous-génome A du blé hexaploïde, et le maïs. Les transitions de couleurs sur le long des chromosomes des génomes en abscisse correspondent à des fusions des chromosomes ancestraux. Les diagonales brisées traduisent des ruptures de la synténie. Quatre évènements remarquables sont matérialisés par des flèches numérotés : (1) exemple de 2 copies d'un protochromosome d'AGK7 issues de la duplication ancestrale ρ, (2) comparaison entre AGK12 et riz indiquant la similarité de leur structure depuis 60 millions d'années, (3) exemple de fusion nichée des protochromosomes 2 et 3 d'AGK12 composant le chromosome 6 du blé, et (4) exemple de 4 copies d'un protochromosome d'AGK7 issues de la combinaison de la duplication ancestrale ρ avec la duplication spécifique du maïs.

4.1.2 Analyse des dynamiques évolutives des SBP

En termes d'évènement évolutif, les SBP correspondent à des sites de réarrangements des chromosomes ancestraux et sont donc le produit d'évènements de fusions, de fissions et de translocations de chromosomes (annexes : Supplementary Figure 3). Sur la base de la liste des 133 SBP identifiés, il est possible de proposer un modèle évolutif en appliquant une approche de parcimonie. Cette approche consiste à considérer que le scénario évolutif le plus probable est celui qui implique le plus petit nombre de réarrangements génomiques (fusions, fissions et translocations) pour expliquer la transition du caryotype ancestral à celui des espèces modernes. Ce scénario (Figure 31) suggère que AGK7 ait été dupliqué pour produire un génome intermédiaire à 14 chromosomes qui, lui-même, subit 2 translocations et 2 fusions de chromosomes pour aboutir à l'ancêtre à 12 chromosomes AGK12 qui constitue le plus récent ancêtre commun aux Poaceae modernes. A partir d'AGK12, les scénarios évolutifs varient radicalement au sein du panel. Le génome du riz conserve la même structure que AGK12, ce qui se traduit l'absence de SBP. Les *Panicoideae*, sorgho, *Setaria* et maïs, comptent respectivement 4, 9 et 38 SBP. Les trois espèces partagent deux fusions de chromosomes nichées, *nested chromosomes fusion* (NCF), contribuant à quatre SBP chez *Setaria* et chez le sorgho et huit SBP chez le maïs. En effet, ces fusions étant communes à l'ensemble de la sous-famille, l'hypothèse de parcimonie implique qu'elles soient le produit d'une unique fusion survenue avant la WGD spécifique du maïs, il y a 5 millions d'années. Le sorgho ne subit par la suite aucun autre réarrangement chromosomique. *Setaria* présente une fusion nichée supplémentaire ajoutant deux SBP et une translocation complexe entre les chromosomes 3 et 7 impliquant une région présentant une fusion nichée antérieure, ajoutant 3 SBP. Le génome du maïs est largement remodelé après la WGD par 16 fusions nichées et une translocation ; il comprend au total 38 SBP. *Brachypodium* compte 14 SBP correspondant à 7 fusions nichées, dont une est partagée avec les *Triticeae* (fusions entre les chromosomes 9 et 12 d'AGK12) et 6 qui sont spécifiques à l'espèce. Les *Triticeae*, orge et blés, présentent 4 fusions nichées, une fusion entre extrémités télomèriques et une fission impliquant le chromosome issu de la précédente fusion AGK(9-12) et les chromosomes AGK3 et AGK11 dans une interaction complexe qui a modelé les chromosomes 4 et 5. L'orge et les génomes de blé diploïdes issus de la spéciation A, B, D il y a 6,5 millions d'années,

ont conservé ce caryotype commun composé de 10 SBP. Le blé tétraploïde et le blé hexaploïde présentent en sus la translocation 4A/7B bien identifiée dans la bibliographie et possèdent donc tous deux 11 SBP.

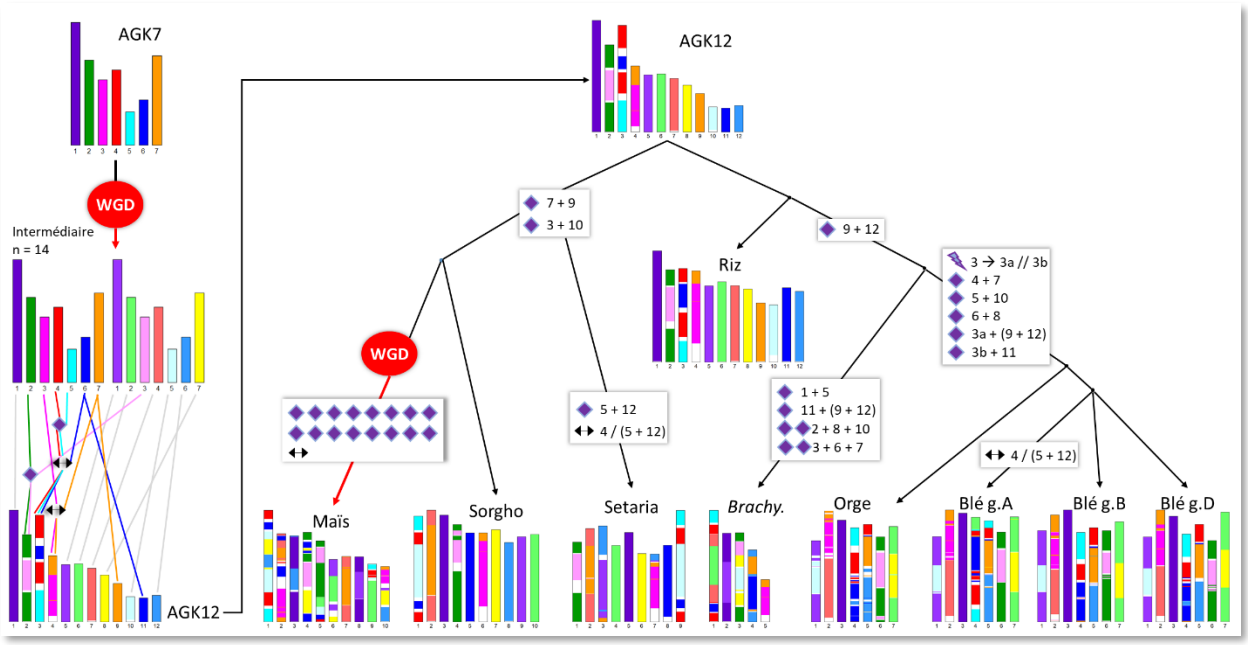


Figure 31 : Scénario évolutif des graminées. Les réarrangements chromosomiques sont représentés par trois symboles : ↗ fission, ◆ fusion, ↔ translocation. Le nombre de symbole correspond au nombre d'occurrence de l'évènement. A gauche : un ancêtre à 7 protochromosomes (AGK7) a subi la WGD ρ pour donner un intermédiaire à 14 chromosomes. Les couleurs des chromosomes de cet intermédiaire dérivent de celles des chromosomes d'AGK7 et sont projetées sur les caryotypes descendants dont ceux des espèces modernes. Cet intermédiaire évolue à travers deux fissions et quatre fusions pour former AGK12 : ancêtre commun le plus récent des graminées qui présente 12 chromosomes. A droite : les différentes spéciations à partir d'AGK12. Pour chaque branche la nature et le nombre de réarrangements sont représentés.

Je me suis ensuite intéressé à la structure des SBP pour tenter de comprendre les mécanismes à l'origine de ces réarrangements chromosomiques. J'ai considéré le blé hexaploïde afin d'évaluer la taille des SBP, c'est-à-dire la taille des régions chromosomiques constituées de mosaïques de gènes issus de deux chromosomes ancestraux. Pour cela j'ai comparé l'ordre des protogènes de AGK12 avec l'ordre des gènes orthologues du blé hexaploïde moderne (Figure 32A).

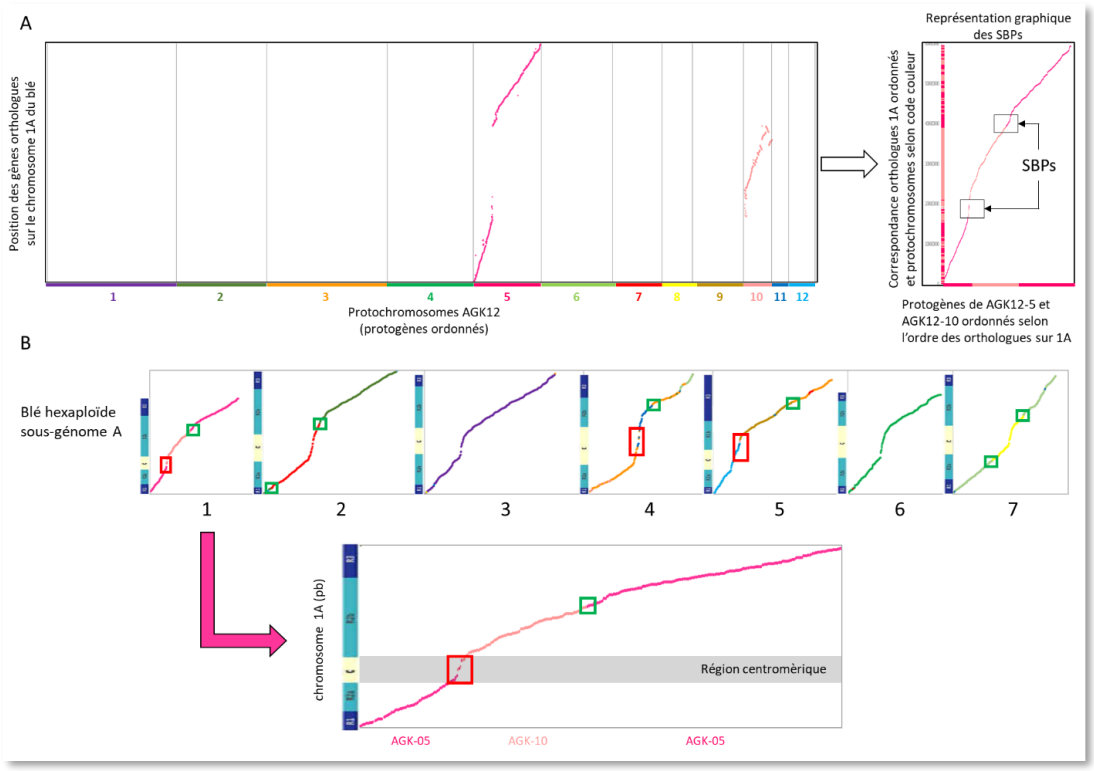


Figure 32 : A. Le chromosome 1A du blé hexaploïde résulte de la fusion des protochromosomes 5 et 10 de AGK12. Cette fusion nichée se caractérise par l'insertion d'un premier chromosome dans un second, suite à une fission au niveau du centromère du dernier et une recombinaison entre les fragments centromériques générés et les extrémités télomériques du premier. Ce réarrangement structural génère 2 SBP et est détecté en observant l'origine ancestrale (représentée par le code couleur des protochromosomes) des gènes de l'espèce moderne. La partie de droite illustre une représentation schématique des SBP. B. Représentation de l'ensemble des points de fusions des chromosomes ancestraux du génome A du blé hexaploïde. En abscisse, représentation des compartiments chromosomiques : télomère (bleu foncé), région péricentromèrique (bleu) et centromère (jaune) (données Appels et al, 2018). Les SBP étendus (cadres rouges) sont localisés dans les centromères ; Les SBP compacts (cadres verts) sont localisés hors des centromères.

Les trois sous-génomes A, B et D présentent les mêmes SBP à l'exception des 2 SBP issus de la translocation 4A/7B. Les 10 SBP communs correspondent aux 4 fusions nichées portées par les chromosomes 1, 2, 4 et 7, et aux deux fusions télomèriques du chromosome 5 (annexes : Supplementary Table 1). Parmi ces 10 SBP, 7 consistent en des régions relativement réduites (quelques centaines de kb) avec des transitions nettes impliquant de faibles nombres de gènes issus de l'un et l'autre des chromosomes ancestraux. A l'inverse, 3 SBP présents sur les chromosomes 1, 4 et 5 couvrent de grandes régions génomiques (plusieurs Mb) et impliquent des nombres conséquents de gènes. En cherchant l'origine de ces différences, j'ai mis en évidence que ces SBP particuliers sont localisés dans les régions centromèriques à la différence des 7 autres SBP localisés dans des régions paracentromèrique ou télomèriques (Figure 32B). L'hypothèse

pour expliquer ces différences est que, lors des fusions chromosomiques qui leurs ont donné naissance, ces SBP étaient des régions de tailles réduites puis ont ensuite subi un brassage des gènes flanquants issus de l'un et l'autre chromosomes ancestraux dans le centromère.

4.1.3 Analyse des dynamiques évolutives des inversions

Pour explorer plus en détail les dynamiques de variations structurales à l'œuvre pendant l'évolution des céréales, en complément des fusions, fissions et translocations qui sont fréquemment caractérisées et étudiées précédemment, j'ai analysé spécifiquement les inversions. La stratégie pour les détecter a consisté à comparer l'ordre des protogènes d'AGK12 à celui des gènes orthologues au sein des espèces modernes. Cette comparaison, dont les résultats sont directement observables en utilisant une représentation type dotplot, a été utilisée pour détecter automatiquement les inversions en analysant les positions relatives des gènes entre l'ordre ancestral et l'ordre moderne. Chaque gène de l'espèce moderne a ainsi été étiqueté, soit non inversé, soit inversé selon qu'il suive ou non l'ordre ancestral. En suivant cette méthode, une inversion est détectée lorsque deux gènes consécutifs sont inversés par rapport à l'ordre ancestral. Afin de se concentrer sur les grandes inversions et pour se prémunir contre des erreurs d'assemblage locales des génomes, j'ai fixé un seuil minimum de 10 gènes modernes inversés. Pour déterminer l'origine évolutive d'une inversion, j'ai comparé les catalogues d'inversions de chaque espèce pour identifier les inversions présentant des gènes en commun. Etant donné que les gènes ancestraux ont pu être perdus différentiellement entre les différentes espèces, il n'y avait pas de sens à chercher des inversions strictement identiques impliquant les mêmes listes de gènes ancestraux. C'est pourquoi les inversions identifiées pour l'ensemble des espèces du panel ont été comparées les unes aux autres, en calculant pour chaque paire d'inversions son indice de Jaccard. L'indice de similarité de Jaccard permet d'évaluer la similarité entre deux ensembles, en l'occurrence les deux ensembles de gènes ancestraux définissants deux inversions à comparer. Deux inversions sont considérées comme ayant une origine évolutive commune, et donc issues du même évènement, lorsqu'elles ont un indice de Jaccard égal ou supérieur à 0,6. Elles constituent alors une famille d'inversion. Sur la base des familles ainsi définies, les inversions ont été analysées une seconde fois pour être fusionnées lorsque deux inversions distinctes mais contiguës au sein d'une même espèce étaient incluses dans une même famille. Ceci afin de

prendre en compte l'ascendance commune des inversions au sein des deux espèces. Après la fusion des inversions, une nouvelle étape de regroupement des familles d'inversions a été réalisée en fonction de leur indice de Jaccard. J'ai ensuite comparé la distribution des inversions entre espèces modernes afin d'inférer leurs origines évolutives. Pour chaque nœud de l'arbre phylogénique, j'ai déterminé le nombre d'inversions partagées par les espèces de la descendance et le nombre d'inversions spécifiques pour chaque espèce moderne. Pour les blés tétraploïdes et hexaploïdes, j'ai considéré individuellement les sous-génomes homéologues A, B et D (Figure 33).

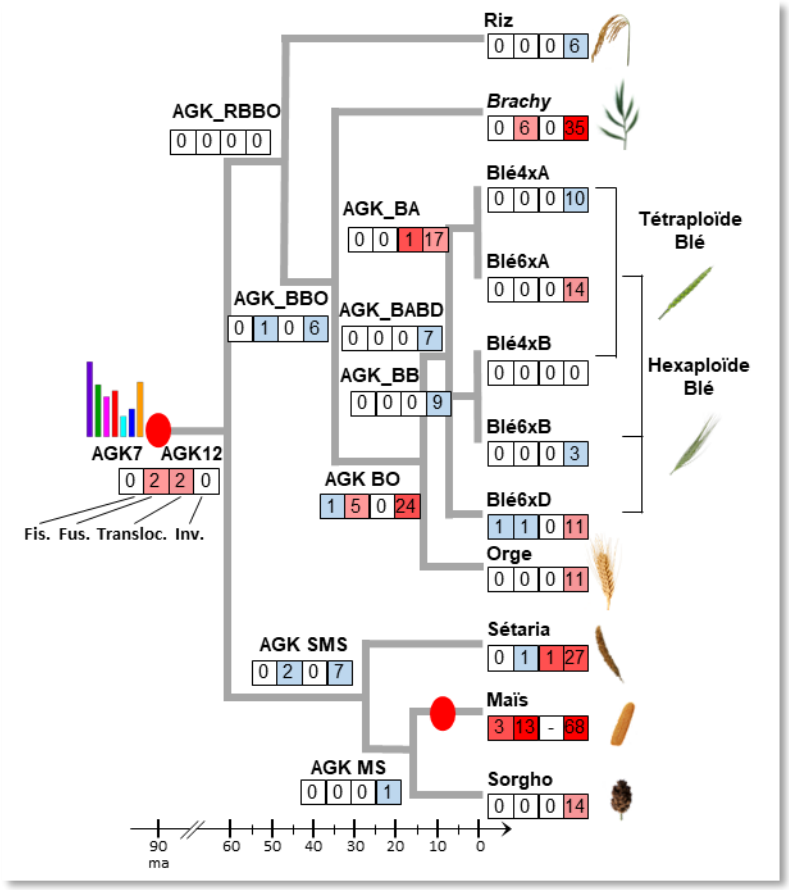


Figure 33 : Représentation de la dynamique des réarrangements chromosomiques chez les graminées depuis AGK7. Les WGD sont représentées par un point rouge. Les boîtes indiquent les nombres de réarrangements : fission, fusion, translocations et inversions. Le code couleur (blanc, bleu clair, rose, rouge, rouge foncé) indique dans cet ordre l'intensité croissante des phénomènes relativement aux autres espèces, sous-génome et reconstruction de caryotypes ancestraux. La flèche graduée permet de dater les évènements en millions d'années.

Le riz est l'espèce ayant subi le moins d'inversions (6) ce qui apparaît cohérent avec le fait qu'il ait conservé la structure AGK originale à 12 chromosomes. Le sorgho présente 22 inversions dont 14 inversions spécifiques. Ces chiffres sont sensiblement différents de ceux du maïs qui présente

76 inversions. Les deux espèces ont 8 inversions en commun, intervenues avant leur divergence il y a 16 millions d'années. Après la divergence, le maïs subi 68 inversions spécifiques qui sont la conséquence du remaniement du caryotype post-polyploïdie. *Setaria* présente 7 inversions communes avec le maïs et le sorgho et 27 inversions spécifiques. *Brachypodium* cumule 41 inversions dont 6 communes avec les *Triticeae* et 35 spécifiques. L'orge présente 41 inversions dont 24 sont communes avec les sous-génomes de blé. Les sous-génomes de blé tétraploïde, A et B, présentent respectivement 64 et 46 inversions. Les sous-génomes de blé hexaploïde, A, B et D, présentent respectivement 68, 49 et 48 inversions. Parmi ces inversions, 37 sont communes à l'ensemble des sous-génomes ; dont 7 spécifiques des blés, 24 partagées avec l'orge et 6 communes avec l'orge et *Brachypodium*. Au sein des sous-génomes, le sous-génome A apparaît comme le plus dynamique à la fois dans le contexte pré-spéciation tétraploïde/hexaploïde avec 17 inversions mais également post-spéciation dans les contextes polyploïdes où le sous-génome A du blé tétraploïde présente 10 inversions et le sous génome A du blé hexaploïde présente 14 inversions. Le sous-génome B présente 9 inversions pré-spéciation et respectivement aucune et trois inversions dans les contextes tétraploïde et hexaploïde. Le sous génome D présente 11 inversions spécifiques, ce qui correspond au total cumulé de 48 inversions, chiffre sensiblement identique à ceux des sous-génomes B. Globalement, il semble que les nombres d'inversions sont corrélés à la dynamique globale des réarrangements chromosomiques, fusions et translocations. Le rapport entre le nombre d'inversion et le nombre cumulé de fusions et de translocation apparaît proportionnel pour l'ensemble des espèces à l'exception des sous-génomes de blé pour lesquels les inversions sont surreprésentées. En effet ceux-ci présentent davantage d'inversions que l'orge ou *Setaria*, deux espèces ayant des nombres de réarrangements chromosomiques comparables (Figure 34).

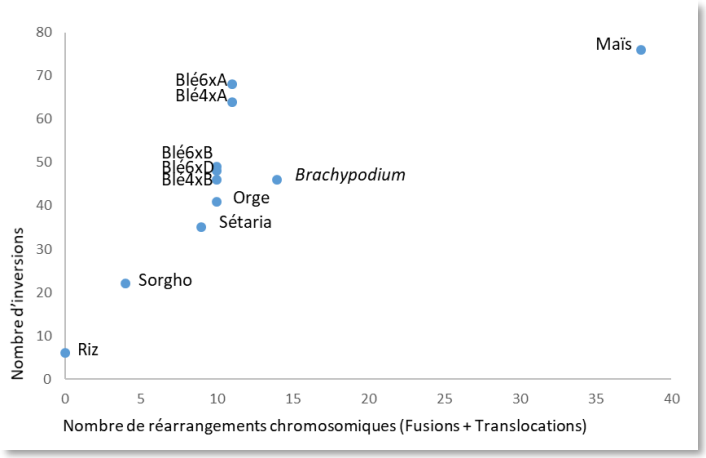


Figure 34 : Rapport entre le nombre de réarrangements chromosomiques (fusions et translocations cumulées) et le nombre d'inversions. Les sous-génomes du blé dans les contextes tétra- et hexaploïde présentent proportionnellement aux nombres de réarrangements chromosomiques davantage d'inversions que les génomes des autres espèces considérées. Dans ce panel, le maïs est l'espèce qui cumule le plus de réarrangements chromosomiques des trois types.

L'observation de la proportion d'inversions (cumul de la taille physique des inversions rapportée à la taille totale du génome), confirme la faible part des inversions chez le riz (0,2%) et leur importance chez les *Triticeae*, notamment chez le sous-génome B du blé tétraploïde (22,8%). *Brachypodium* présente un rapport de 10,7%, tandis que les *Panicoideae* affichent des chiffres compris entre 12,6% et 15,6% %. Les *Triticeae* présentent les proportions d'inversions les plus élevées du panel, 17,4% pour l'orge et de 19,4% à 22,8% pour les sous-génomes du blé tétraploïde et hexaploïde (annexes : Supplementary Figure 4b).

J'ai calculé la taille moyenne et la densité en gènes des inversions. J'ai utilisé ces métriques pour déterminer si les inversions diffèrent, en termes de structure, du reste du génome. Pour cela, pour chaque espèce, 1000 simulations, consistant à distribuer aléatoirement sur les génomes des inversions sur la base du nombre et de la taille des inversions observées pour chaque espèce, ont été réalisées et la densité de gènes (en nombre de gènes par Mb) à l'intérieur des inversions simulées a ensuite été calculée. Pour toutes les espèces, la densité en gènes observée dans les inversions observées était supérieure à la majorité des valeurs obtenues dans les simulations. *Brachypodium* présente une densité de gènes observée supérieure à 87,3% des valeurs simulées tandis que cette valeur est supérieure à 97,9% pour toutes les autres espèces du panel (Annexes : Supplementary Figure 4d). La densité en gènes élevée dans les inversions par rapport au reste du génome pourrait refléter leur localisation télomèrique ; les télomères étant des compartiments

chromosomiques connus pour présenter une forte densité en gènes. Pour évaluer cette hypothèse, j'ai divisé les génomes en fonction du taux de recombinaison le long des chromosomes en deux compartiments, respectivement LR (*Low Recombination*) correspondant aux centromères et péri-centromères et HR (*High Recombination*) correspondant aux télomères (table en annexe). Les compartiments LR et HR sont corrélés à la densité de gènes, respectivement de faible à élevée. Pour chaque espèce, les inversions ont été assignées aux compartiments LR et HR. Les proportions des nombres d'inversions dans chaque compartiment ont été comparées aux fractions de l'espace physique que ces compartiments occupent respectivement sur les chromosomes (à l'exclusion du riz en raison du très faible nombre d'inversions). Pour toutes les espèces étudiées, la proportion d'inversions située dans le compartiment LR est inférieure à la taille physique relative du compartiment HR (annexes : Supplementary Table 3) ce qui indique que les inversions ne sont pas distribuées uniformément le long des chromosomes et tendent à être localisées au niveau des télomères, zones de forte recombinaison et de forte densité en gènes.

Au-delà des gènes qui sont des régions conservées au cours du temps permettant de reconstruire l'histoire évolutive, j'ai considéré l'hypothèse que les régions impliquées dans les réarrangements chromosomiques sont des régions riches en éléments répétés et notamment en ET. Pour étudier cet aspect, qui est par construction absent des analyses basées sur la reconstruction des ordres des gènes des caryotypes ancestraux, j'ai étudié des inversions identifiées entre des chromosomes homologues des génomes tétraploïde et hexaploïde de blé. J'ai ainsi comparé la séquence du blé hexaploïde *Triticum aestivum cv. Chinese spring* (Appels *et al.*, 2018) à celle de deux variétés de blé tétraploïde, une variété sauvage *Triticum turgidum ssp. Dicoccoides cv. Zavitan* (Avni *et al.*, 2017) et une variété cultivée *Triticum turgidum ssp. durum cv Svevo* (Maccaferri *et al.*, 2019). J'ai identifié dans chaque cas une inversion et je me suis appuyé sur l'annotation du génome du blé hexaploïde (https://urgi.versailles.inrae.fr/jbrowseiwgsc/gmod_jbrowse/) pour connaître la nature des séquences de l'inversion (Figure 35).

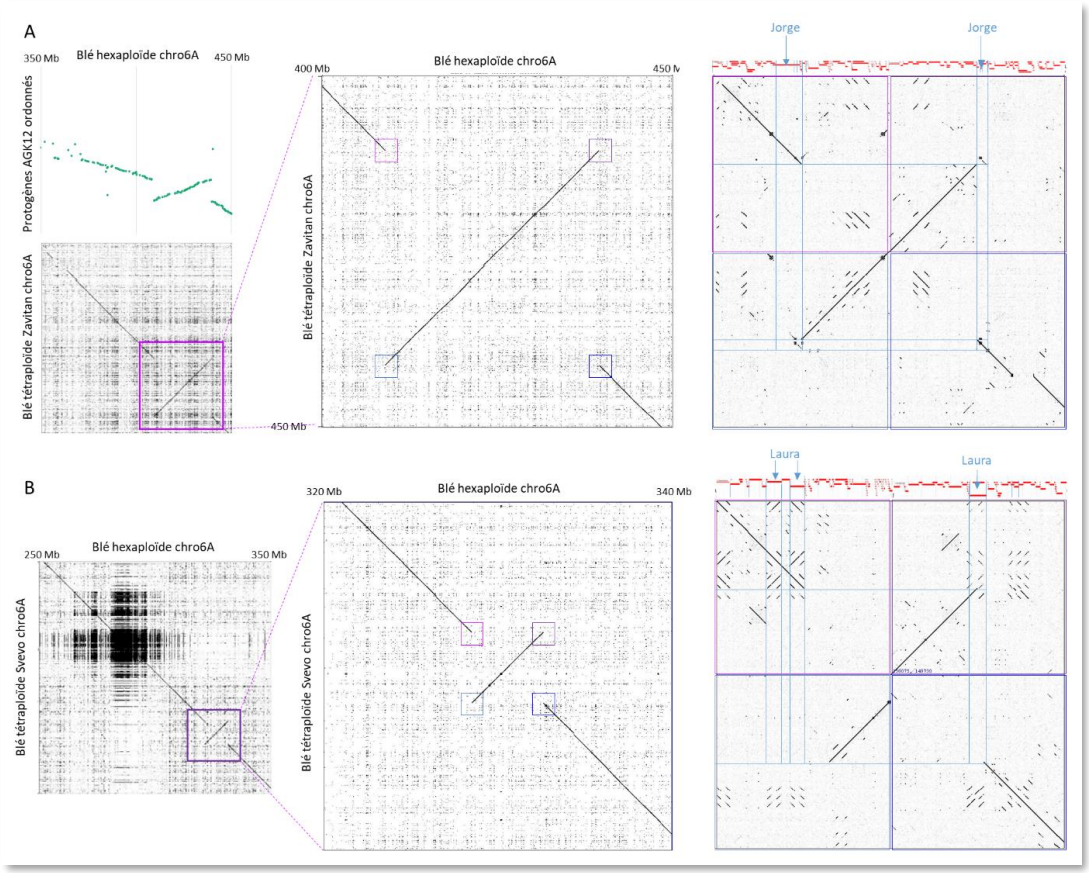


Figure 35 : Dotplot présentant deux inversions observées sur les chromosomes homologues 6A entre le blé hexaploïde et les blés tétraploïdes Zavitan (AB) et Svevo (B). De gauche à droite dotplot 100Mb, dotplot de la région inversée et dotplot entre des bornes de l'inversion au regard de l'annotation des éléments répétés sur le blé hexaploïde Chinese spring.

Dans les deux cas étudiés des ET de même natures se trouvent sur les séquences flanquant les inversions. Ces éléments transposables sont de type Copia (Jorge) dans le cas de l'inversion du génotype Zavitan et de type Gypsy (Laura) dans le cas de l'inversion du génotype Svevo. Bien que très parcellaires ces constatations accréditent l'hypothèse que les inversions résultent d'homologies de séquences entre ET occasionnant des recombinaisons illégitimes.

4.1.4 Analyse des dynamiques de mutation des gènes post-polyploïdie

Après l'étude de l'évolution des structures chromosomiques des génomes, la dynamique évolutive des gènes a été examinée. Les 447 744 gènes annotés sur l'ensemble des espèces du panel ont été utilisés pour détecter les relations d'homologie par une analyse d'alignements multiples et de construction d'arbres phylogénétiques par gènes réalisés par Wandrille Duchemin

et dont j'ai exploité les résultats pour l'analyse des dynamiques des gènes comme décrit dans la suite du manuscrit. Cet alignement a permis d'identifier 304 549 gènes présentant des relations d'homologie, distribués en 15 810 familles. Parmi ces 15 810 familles, 2 599 et 3 851 correspondent respectivement à des gènes AGK7 (pré-ρ) et AGK12 (post-ρ), conservés dans toutes les espèces étudiées. Ces familles de gènes ont permis d'étudier les variations de taux de mutation au sein des gènes en fonction de leur histoire évolutive. Pour cela, j'ai observé les changements de taux de substitution au cours de l'évolution des céréales, exprimés en nombre de substitutions par site par milliard d'années. Ces taux ont été calculés pour chaque nœud de l'arbre évolutif. La WGD ancienne (ρ) et la WGD récente du maïs sont toutes deux associées à un nombre accru de substitutions (respectivement 4,11 et 9,79), respectivement 4^{ième} et 1^{ere} valeurs parmi l'ensemble des valeurs associées aux branches de l'arbre évolutif. Après la WGD ancestrale ρ, les taux d'évolution ralentissent nettement dans les deux lignées qui se différencient à partir d'AGK12 (2,15 pour les *Pooideae* et 2,04 pour les *Panicoideae*). Dans les deux sous-familles, les branches menant à des évènements de spéciations (AGK-BBO, AGK-BB, AGK-MS, nommées d'après les initiales des espèces modernes) présentent des taux de substitutions, 5,04, 3,23 et 4,3 respectivement, plus élevés que ceux des branches conduisant à une espèce (2,74 pour le riz, 2,17 pour *Brachypodium*, 3,07 pour l'orge, 2,84 ou moins pour les blés, 2,31 pour Setaria, 1,93 pour le sorgho) (Figure 36).

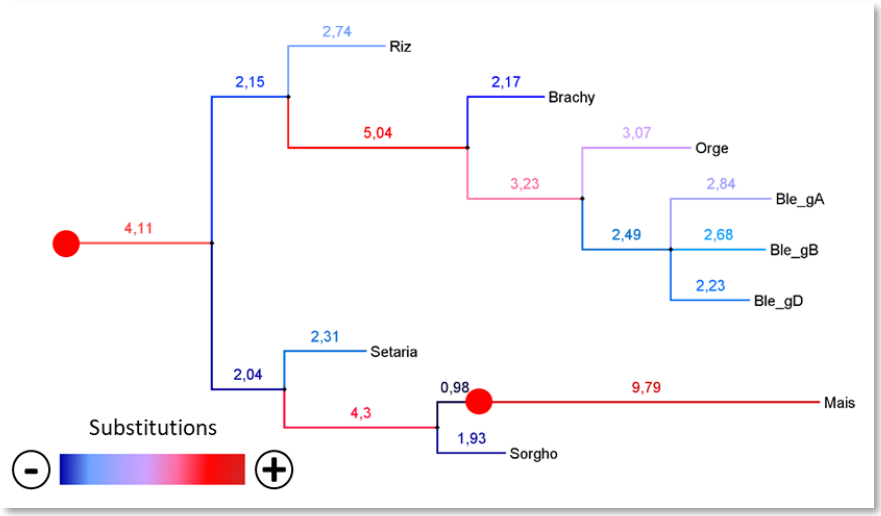


Figure 36 : Taux de substitution en fonction du temps (nombre de substitutions par site par milliard d'années) pour les 7 espèces du panel incluant les 3 sous-génomes du blé pré-polyploïdisation. Les évènements de polyploïdisation sont marqués par les points rouges. La dynamique des substitutions est proportionnelle à la longueur des branches et matérialisée par le dégradé de bleu à rouge.

Pour mieux comprendre l'évolution des taux de substitutions dans la lignée des *Triticeae*, j'ai considéré un ensemble de 3 905 gènes orthologues conservés dans chaque génome ou sous-génomes pour les espèces polyploïdes comprenant l'orge, les blés diploïdes *Triticum urartu* et *Aegilops tauschii*, le blé tétraploïde et le blé hexaploïde (Figure 37). Ce set de gènes spécifique des *Triticeae* correspond à 13 millions d'années d'évolution et permet d'analyser la divergence récente des sous-genres de blé, en comparant les taux de substitutions entre les différentes lignées.

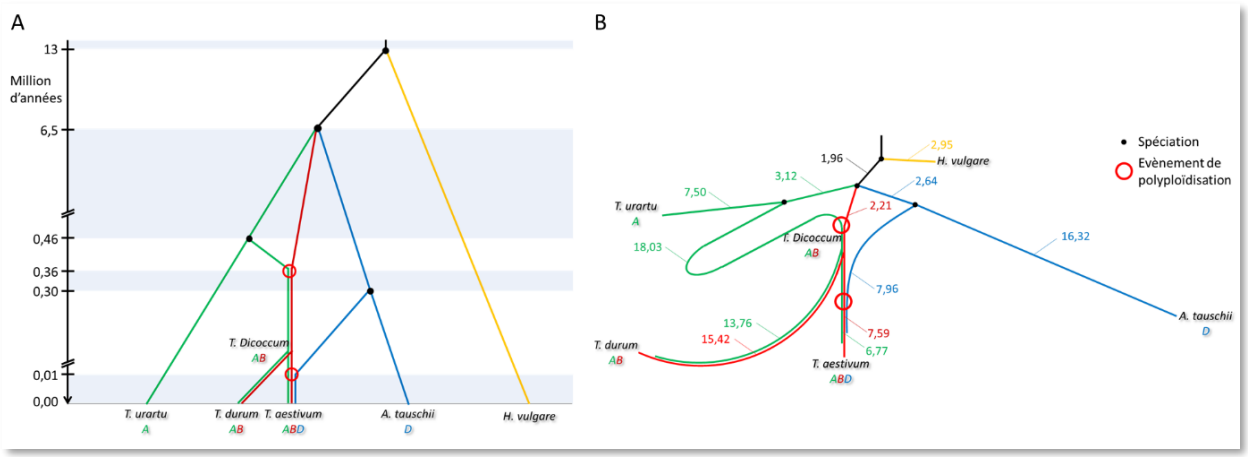


Figure 37 : A. Arbre phylogénétique des Triticeae. Les longueurs des branches représentent les durées écoulées (en millions d'années) depuis les événements de spéciation (points noirs) et de polyploïdisation (cercles rouges) B. Variations de la dynamique des substitutions dans la lignée des Triticeae. Les longueurs des branches sont proportionnelles aux taux de substitution exprimés en nombre de substitutions par site par milliard d'années, mettant en évidence les différences de dynamique des taux de substitution entre espèces et sous-génomes.

La lignée de l'orge et les lignées de blé avant les événements de spéciation et de polyploïdisation présentent des taux de substitution faibles, compris entre 1,96 à 3,12. Au sein des blés, le compartiment A apparaît très dynamique à partir de la divergence avec *Triticum urartu*, il y a 460 000 ans, dans le contexte diploïde (taux de 18,03, le plus élevé observé) et le contexte tétraploïde (13,76). Dans le contexte hexaploïde, le génome A (6,77) évolue légèrement moins vite que *Triticum urartu* (7,59). Pour la lignée D, cette dynamique est inversée puisque l'évolution du sous-génome D du blé hexaploïde est plus lente (7,96) que celle de son homologue diploïde sauvage *Aegilops tauschii* (16,32). La nette différence dans l'historique des mutations entre les blés tétraploïdes et hexaploïdes pour les sous-génomes A (6,77 chez l'hexaploïde et 13,76 chez le tétraploïde, correspondant à un rapport de 49%) et B (7,59 chez l'hexaploïde et 15,42 chez le tétraploïde, pour un rapport de 49%) peut illustrer à la fois l'effet de l'événement de

polyploïdisation additionnel, mais aussi d'autres processus comme la domestication et la sélection. Les rapports observés entre les taux de substitution des sous-génomes tétraploïdes et hexaploïdes sont indépendants des dates de divergence choisies étant donné que les deux taux se rapportent au même ancêtre commun, traduisant par conséquent l'accumulation des substitutions accumulées depuis cette étape de spéciation.

4.1.5 Conclusions

- **Les génomes des graminées ont évolué à partir d'un ancêtre de 7 chromosomes avec 16 000 gènes conservés ancestralement. Ces gènes correspondent à des fonctions cellulaires de base. Les génomes ont évolué par le biais de fusions et de fissions chromosomiques sans que l'on puisse détecter de réutilisation des SBP au cours du temps ou dans des branches distinctes. Les événements d'inversion ont tendance à se produire au niveau des télomères et plus fréquemment dans un contexte de polyploïdie que chez les diploïdes.**

- L'étude des dynamiques évolutives des gènes, en observant les taux de substitutions au sein des gènes ancestraux conservés par l'ensemble des espèces, montre une accumulation de mutations des séquences avant les évènements de spéciation. Chez le blé, il est notable que les sous-génomes tétraploïdes accumulent plus de mutations que leurs homologues hexaploïdes.

4.2 Impact de la polyploïdie sur la *structure* des gènes

4.2.1 Les pertes de gènes façonnent les génomes des polyploïdes

Définition des compartiments LF et MF

De nombreuses études (voir page 61) proposent que la WGD ancestrale (ρ) soit suivie d'un phénomène de dominance des sous-génomes. Ce phénomène résulte de pertes de gènes biaisées entre les génomes post-polyploïdie, ou fractionnement asymétrique, qui définissent des compartiments LF (*Least Fractionated*) et MF (*Most Fractionated*) le long des chromosomes des espèces modernes. Dans cette étude, il était nécessaire d'inférer les compartiments LF et MF de

façon homogène pour toutes les espèces du panel. Pour cela, les reconstructions des caryotypes ancestraux AGK (AGK7 et AGK12) ont été utilisés et une approche statistique basée sur le comptage et la comparaison du nombre de gènes ancestraux retenus dans chacun des deux compartiments chromosomiques issus des WGD a été développée. La WGD ancestrale ρ, les WGD spécifiques du maïs et celles des blés tétraploïde et hexaploïde ont ainsi été examinées. Pour ce faire, des fenêtres glissantes de 100 gènes ancestraux consécutifs sont considérées. Pour chaque fenêtre les pertes de gènes dans les deux régions homéologues issues d'une WGD sont comptabilisées (Figure 38). S'il existe une différence significative entre les deux compartiments, celui ayant conservé le plus de gènes est noté LF et le compartiment homologue ayant perdu le plus de gènes est noté MF. En fonction de leur appartenance à l'un ou l'autre compartiment, les gènes se voient attribuer une annotation LF et MF ; on parle donc de gènes LF et de gènes MF.

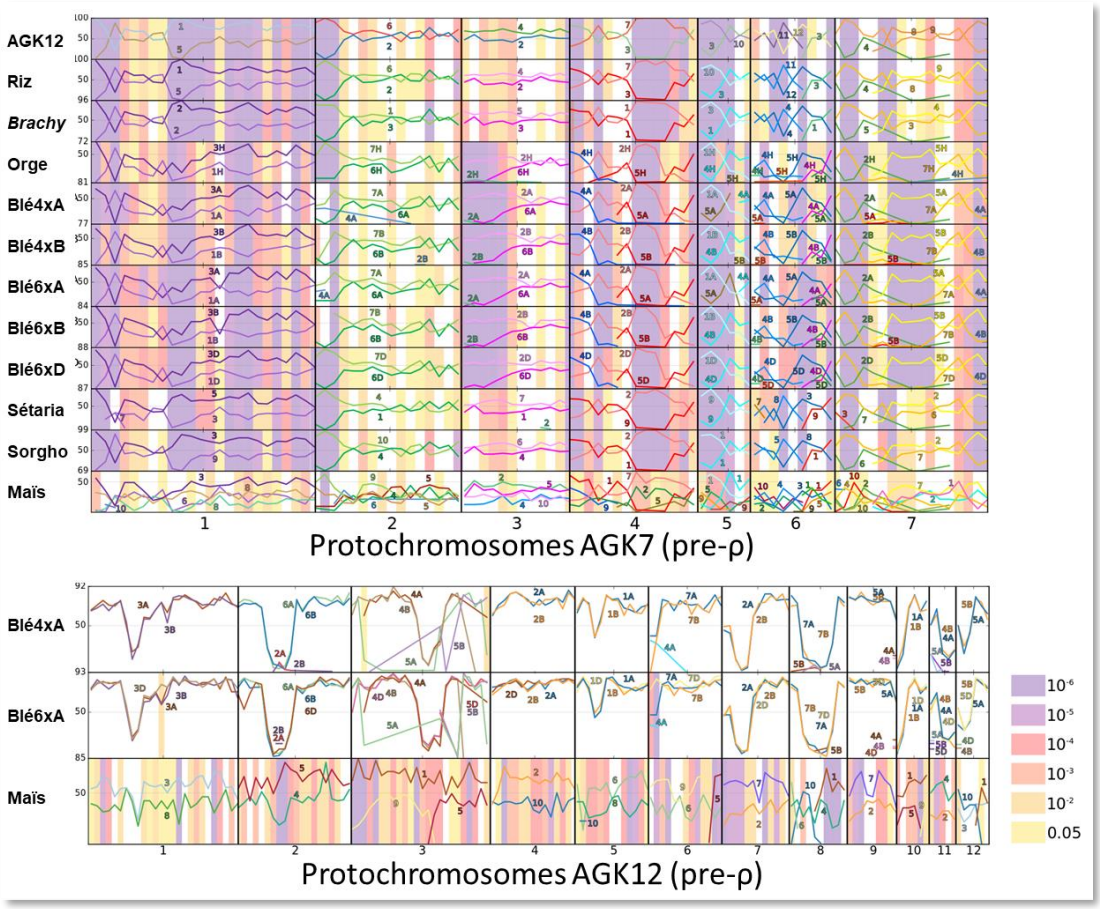


Figure 38 : Inférence des compartiments LF/MF en dénombrant dans des fenêtres de 100 protogènes ancestraux ordonnés le nombre de gènes ancestraux conservés chez les espèces modernes sur chacun des deux segments chromosomiques issus d'un évènement de WGD. En abscisse, les protogènes ancestraux ordonnées sur les sept

protochromosomes d'AGK7 (panneau supérieur) et d'AGK12 (panneau inférieur). En ordonnée : les génomes et sous génomes des espèces modernes issues de la WGD d'AGK7 (panneau supérieur) ou des WGD d'AGK12 spécifiques des blés et du maïs (panneau inférieur). Pour chaque espèce ou sous-génome, deux courbes sont numérotées en fonction du numéro de chromosome moderne correspondant au protochromosome d'AGK7. Les courbes suivent les nombres de protogènes ancestraux conservés sur chacun des segments homéologues par fenêtre de 100 protogènes. Les couleurs de fond correspondent à la valeur de p-value issue de la comparaison entre les deux courbes pour une fenêtre donnée est significative. Si la valeur de p-value est inférieure à 0,05, la différence est significative : le segment chromosomique ayant conservé le plus de gènes est étiqueté LF et le segment homéologue ayant perdu le plus de gènes est étiqueté MF.

Sur la base de la WGD ancestrale ρ à partir de l'ancêtre AGK7, cette approche a permis de définir, au sein de l'ensemble des génomes du panel, les statuts de 158 124, 74 311 et 215 309 gènes, respectivement, LF, MF et non significatif. Sur la base de la WGD spécifique du maïs, à partir de l'ancêtre AGK12 29 148, 17 455 et 16 877 gènes ont été annotés, respectivement, LF, MF ou non significatif (Tableau 6.A).

Définition du statut des gènes ancestraux : conservés en paire ou singletons

Outre l'appartenance des gènes aux compartiments LF, MF ou ne présentant pas de biais, j'ai observé les impacts subis par les gènes homéologues post-WGD, selon qu'ils soient conservés en paire ou en copie unique, notée singleton, après la perte d'une des copies du gène ancestral dupliqué (Tableau 6.B). Si la définition du statut en paire ou singleton semble triviale, il était cependant important de la définir précisément pour pouvoir l'attribuer de façon standardisée et automatisée au sein des espèces du panel. Une première façon de répertorier le statut des gènes ancestralement dupliqués consiste à considérer les 10 286 protogènes de AGK7 et compter combien parmi les gènes descendants chez les espèces modernes sont conservés en paires ou se trouvent en copie unique. Notée « AGK7-derived », cette méthode permet d'inférer 75 662 singletons et 19 032 gènes conservés en paires au sein des espèces étudiées. Elle présente cependant deux inconvénients. D'une part, elle ne permet de définir le statut de seulement 21% des 447 744 gènes du panel et, d'autre part, elle ne prend pas en compte des gènes ayant réellement une origine ancestrale pré-ρ mais qui n'ont pas pu être retenus dans la reconstruction, car absents dans au moins deux des trois espèces utilisées pour la reconstruction des caryotypes ancestraux. Pour prendre ces gènes en compte, une seconde méthode, dite « AGK12-derived » a été utilisée. Les singletons sont définis comme les gènes ayant une copie ancestrale dans l'ancêtre AGK12 et qui se retrouvent en copies uniques dans les espèces modernes. Etant présent dans

AGK12, ils sont le produit de la WGD ρ et ont existé sous forme de paires de gènes homéologues, dits ohnologues, puis ont été perdus entre la WGD ancestrale et l'ancêtre le plus récent des céréales, AGK12, ou, par la suite au fil des spéciations. Par complémentarité, sont référencés comme gènes en paires, les gènes des espèces et sous-génomes pour lesquelles deux ohnologues ont été inférés au sein d'une espèce. Cette méthode infère 156 560 singletons et 113 837 gènes conservés en paires soit 60% des gènes du panel.

A

	WGD ρ	WGD maïs
LF	158124|35%	29148|46%
MF	74311|17%	17455|28%
N/A	215309|48%	16877|26%

B

	AGK7-derived	AGK12-derived
Singletons	75662|17%	156560|35%
Gènes en paires	19032|4%	113837|25%
N/A	353050|79%	177347|40%

C

WGD ρ x AGK12-derived	LF	MF	N/A
Singletons	63404|14%	27067|6%	66089|15%
Gènes en paires	44317|10%	22600|5%	46920|10%
N/A	50403|11%	24644|6%	102300|23%

Tableau 6 : Effectifs des gènes en fonction des différents compartiments et statuts, des WGD considérés (pour LF/MF), des méthodes d'inférence (pour singleton et gène en paire). A : effectifs de gènes LF/MF issus de la WGD ρ et de la WGD spécifique du maïs. B : effectifs de gènes singleton et en paire selon les deux méthodes d'inférence. C : effectifs de gènes selon le double statut LF/MF et singleton/paire. N/A correspond aux effectifs de gènes pour lesquels l'un ou l'autre statut n'est pas défini.

Sur la base de leur appartenance à un compartiment LF ou MF et de leur statut, conservé en paire ou singleton, les gènes peuvent être caractérisés en fonction des deux critères (Tableau 6.C).

4.2.2 Analyse des dynamiques de mutations des gènes selon leurs statuts

Sur la base des compartiments (LF/MF) et des statuts des gènes (Singleton/Pair), j'ai étudié la plasticité des gènes *via* les mesures de Ka, Ks, du rapport Ka/Ks et des longueurs de branches. La mesure du Ka et du Ks se base sur l'alignement de deux gènes homologues. Le Ka correspond à la mesure du nombre de mutations non synonymes par rapport au nombre de sites non synonymes, c'est-à-dire tous changements de la séquence nucléotidique dans un triplet de nucléotide donné qui modifient la nature de l'acide aminé codé par ce triplet. Le Ks correspond à la mesure du nombre de mutations synonymes par rapport au nombre de sites synonymes, c'est-à-dire les changements de nucléotide dans un triplet de nucléotide donné qui ne modifient pas le type d'acide aminé codé (Figure 39). Le rapport Ka/Ks est fréquemment utilisé pour caractériser le type de sélection à l'œuvre. Un rapport Ka/Ks supérieur à 1 traduirait une sélection positive, s'il est égal à un la sélection serait neutre et s'il est inférieur à un une sélection purificatrice opèrerait.

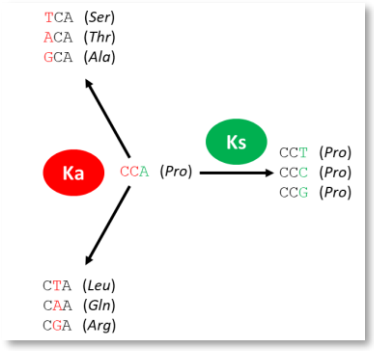


Figure 39 : Différence entre Ka et Ks par un exemple illustrant pour un codon donné l'impact de la mutation d'un des trois nucléotides. La mutation du premier ou du deuxième nucléotide induit un changement de l'acide aminé codé et constitue donc une mutation non synonyme dont la fréquence est mesurée par la valeur de Ka. La mutation du troisième nucléotide ne modifie pas la nature de l'acide aminé codé et constitue donc une mutation synonyme dont la fréquence est mesurée par la valeur de Ks.

Les mesures des longueurs de branches des arbres phylogénétiques de gènes ont également été considérées. Alors que les mesures de Ka et de Ks agrègent la totalité du chemin évolutif entre deux copies homologues, l'information apportée par la mesure de longueur de branche est une mesure de la dynamique évolutive entre deux nœuds de l'arbre (entre deux ancêtres ou entre le dernier ancêtre commun et l'une des espèces modernes).

Ainsi, à chaque position dans l'arbre évolutif des graminées (pour un événement de spéciation ou de duplication), il est possible d'avoir un catalogue de gènes (l'ensemble des gènes conservés entre les espèces descendants d'un évènement de spéciation ou les gènes en paires issus d'un événement de duplication) et de calculer les taux moyens de Ka, Ks et Ka/Ks, ainsi que les longueurs de branche moyennes associés.

L'analyse comparative du taux d'évolution entre les singletons et les gènes en paires révèle certaines différences significatives. Dans le cas des substitutions non synonymes (Ka) aucune tendance claire ne se dégage. 6 des 9 ancêtres considérés ne montrent pas de différence, tandis que les singletons accumulent davantage de substitutions non synonymes que les paires dans AGK12 et AGK-BBO et qu'à l'inverse les paires présentent plus de substitutions que les singletons dans AGK-MS. La tendance est plus marquée dans le cas des substitutions synonymes (Ks) pour lesquelles les paires affichent un nombre significativement plus élevé de substitutions synonymes que les singletons pour 6 des 9 ancêtres considérés, tandis qu'il n'y a pas de différences significatives pour les trois derniers cas. Ces différences se traduisent de la même façon (paires >

singletons) pour le rapport Ka/Ks pour 4 ancêtres sur 9, trois ancêtres ne présentant pas de différences significatives et l'ancêtre commun des blés montrant une tendance inverse (paires < singletons). Les rapports de longueurs de branches montrent deux tendances distinctes. Chez les ancêtres les plus anciens, AGK12 (60mya), AGK-RBBO (46mya) et AGK-BBO (35mya), les singletons ont évolué plus rapidement que les paires, tandis que chez 4 des 6 ancêtres les plus récents, AGK-SMS (27mya), AGK-MS (16mya), AGK-BO (13mya), AGK-BABD (6.5mya) et chez 4 des 11 espèces modernes, les paires ont évolué plus rapidement que les singletons. Les deux ancêtres des sous génomes A et B de blé et 7 espèces modernes ne présentent pas de différences. Ces chiffres de longueurs de branches confirment la tendance générale selon lequel les paires ont évolué plus rapidement que les gènes singletons, principalement dans les ancêtres plus récents (à partir de ~27mya) et les espèces modernes. Globalement, ces résultats suggèrent que la vitesse d'évolution est affectée par le fait que le gène a été conservé en copies multiples (paires) ou non (singletons) après l'événement WGD ρ, les singletons étant alors plus stables en termes d'accumulation de mutations que les paires (Figure 40.A).

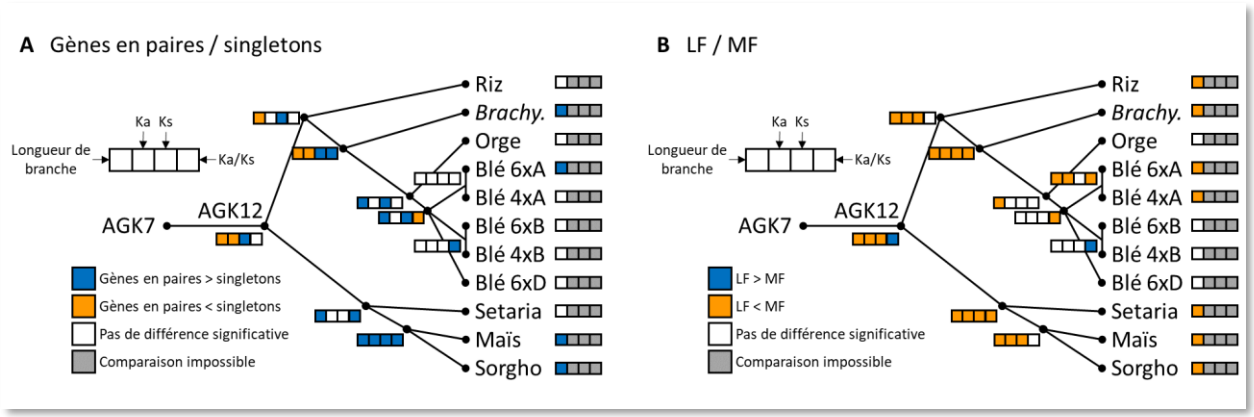


Figure 40 : Comparaison des dynamiques de mutations selon les valeurs de substitution (longueurs de branches), Ka, Ks et rapport Ka/Ks entre les ensembles de gènes (A) conservés en paires ou singletons et (B) étiquetés LF ou MF.

En ce qui concerne les compartiments LF et MF, les quatre métriques (longueur de branches, Ka, Ks, Ka/Ks), appliquées aux gènes selon leur appartenance à l'un ou l'autre compartiment, indiquent que les gènes LF ont globalement évolué plus lentement que les gènes MF. En effet, les gènes LF montrent significativement moins de substitutions synonymes (Ks) que les gènes MF chez cinq des neuf ancêtres, ne montrant aucune différence significative pour les quatre autres, ceux-ci étant les ancêtres les plus récents, c'est-à-dire les plus éloignés de l'événement original ρ

WGD, pour lesquels il est concevable que l'effet devienne trop faible pour être détecté. Des résultats similaires sont trouvés pour les substitutions non synonymes (Ka) chez cinq des neuf ancêtres qui affichent des valeurs plus élevées de substitutions dans les gènes MF. Les quatre cas ne montrant pas de différence significative sont ceux des mêmes ancêtres récents que dans l'analyse de Ks. La comparaison des rapports Ka/Ks présente un schéma plus complexe. Le rapport Ka/Ks est plus faible pour les gènes LF que pour les gènes MF chez trois des neuf ancêtres, sans différences statistiques chez quatre ancêtres (AGK-MS et AGK-RBBO) et significativement plus élevé pour les gènes LF chez deux ancêtres, l'un très ancien (AGK12) l'autre très récent (l'ancêtre des sous génomes B de blé). L'analyse des longueurs de branches vient confirmer les tendances des Ka et Ks. Les gènes de la fraction MF présentent des branches significativement plus longues que celles des gènes LF chez sept des neuf ancêtres considérés. Aucune différence significative n'a été détectée chez l'ancêtre des blés ABD et l'ancêtre des sous génomes B de blé ainsi que pour 4 espèces existantes appartenant aux *Triticeae* (Figure 40.B). L'ensemble des résultats montrent que la vitesse d'évolution des gènes est affectée par leur appartenance à l'une ou l'autre des fractions LF ou MF du génome. Les gènes de la fraction LF sont globalement plus stables que ceux de la fraction MF qui accumulent davantage de mutations. Les différences observées semblent s'amoindrir au cours du temps, plus l'évènement de WGD étant ancien.

Les compartiments LF et MF et le statut en paire ou singleton, étant deux conséquences du même phénomène de pertes de gènes post WGD, sont interconnectés. Les résultats précédents suggèrent que la vitesse d'évolution est affectée à la fois par l'appartenance du gène à la fraction LF ou MF et par le fait que le gène a été conservé en copies multiples (paires) ou non (singletons). En termes de fonction, sur la base des enrichissements en GO, il apparaît qu'il y a une convergence entre les singletons et les gènes appartenant à la fraction LF, d'une part, et les gènes conservés en paires et les gènes appartenant à la fraction MF d'autre part. Les premiers, singletons/LF, présentent une surreprésentation de gènes impliqués dans des processus de signalisation (transducteur, transférase, transporteur, kinase, hydrolase). Les seconds, paires/MF, sont enrichis pour les activités de transcription (liaison, transcription).

4.2.3 Conclusions

L'analyse des dynamiques de pertes de gènes des 8 espèces de céréales constituant le panel d'étude a permis :

- D'inférer des compartiments LF et MF au regard de l'évènement de duplication ancestral ρ pour l'ensemble des espèces et en fonction de la duplication récente pour le maïs,

- De définir le statut des gènes, conservés en paire ou singleton au regard de l'évènement de duplication ancestral ρ.

Sur la base de ces statuts les dynamiques de mutations ont été étudiées et il a été montré que :

- Les séquences des singletons apparaissent majoritairement plus stables que celles des gènes conservés en paires

- Les séquences des gènes LF apparaissent majoritairement plus stables que celles des gènes MF

- Les singletons et les gènes LF sont enrichis pour les mêmes fonctions, à savoir les processus de signalisation, tandis que les gènes conservés en paires et les gènes MF sont enrichis pour les activités de transcription.

4.3 Impact de la polyploïdie sur la régulation des gènes

Trois espèces du panel, *Brachypodium*, le maïs et le blé hexaploïde ont été sélectionnées pour étudier l'impact des WGD sur la régulation des gènes. Ces espèces sont caractérisées pour l'expression des gènes (RNAseq) au cours du développement du grain, de la diversité nucléotidique (SNP) au niveau populationnel et de la méthylation (BSseq) pour *Brachypodium* et le maïs. Ces trois espèces représentent trois trajectoires évolutives distinctes depuis la WGD ρ, et par conséquent, apportent des informations complémentaires dans le cadre de cette étude. *Brachypodium* est un modèle de néodiploïdisation, avec des pertes de gènes importantes et de nombreux réarrangements. Le maïs a subi un nouvel évènement de polyploïdisation, il y a environ 5 millions d'années, suivi d'une nouvelle phase de pertes de gènes et d'intenses réarrangements du caryotype. Le blé présente une hexaploïdisation récente, il y a moins de 500 000 ans, au cours

de laquelle les structures chromosomiques de trois sous-génomes se sont majoritairement maintenues et les pertes de gènes ont été modérées. Il est à noter que pour les gènes du maïs, à la fois pour leur statut, paires ou singletons, et leur appartenance aux fractions LF ou MF, il est possible de distinguer l'impact de la WGD ancestrale (90 mya) de celui de la WGD la plus récente (5 mya). Sur la base de ces données, j'ai exploré les dynamiques d'expression, de méthylation, et de mutations aux échelles interspécifique (entre espèces) et intraspécifique (entre gènes homologues au sein d'une même espèce).

4.3.1 Panorama des données omiques analysées

Pour réaliser le travail d'analyses décrits dans la suite du manuscrit, j'ai travaillé sur la base de données obtenues à partir de la production de matériel végétal par les équipes de Bertrand Dubreucq (*Brachypodium)* et de Peter Rogowsky (maïs), d'ADN et d'ARN par Caroline Pont et Cécile Huneau de l'équipe de Jerome Salse, de séquences par la société INTEGRAGEN. Les séquences obtenues ont été cartographiées sur les génomes de références par Mamadou Dia Sow et David Armisen de l'équipe de Jerome Salse.

Données d'expression

L'expression des gènes dans les trois espèces est étudiée en utilisant des données de séquençage des ARNm (RNAseq) du grain à trois stades concordants du remplissage. Les grains ont été prélevés sur des plants de *Brachypodium* (génotype BD21-3), de maïs (génotype B73) et de blé (génotype Récital) à trois stades de développement considérés comme équivalents sur la base de critères morphologiques, à savoir 9, 16 et 28 jours après pollinisation, *days after pollination* (DAP), pour *Brachypodium*, 7, 15 et 35 DAP pour le maïs et 100, 250 et 500 degrés jours, *degree-day* (°D) pour le blé. Par commodité, dans la suite de ce document les trois stades synchrones seront désignés par les valeurs en degrés jours des trois stades de développement du blé (100, 250, 500 °D). Les ARN totaux sont extraits des grains par la méthode phénol / chloroforme / alcool isoamylique, puis purifiés afin de s'assurer de l'absence de trace d'ADN. Les échantillons d'ARN totaux sont fragmentés puis une hybridation à un oligonucléotide poly(T) est réalisée pour sélectionner les ARNm sur la base de leur queue poly(A), succession caractéristique de nombreuses Adénosine (A) à l'extrémité 3' des ARNm. Les ARNm sélectionnés sont transcrits

en ADN complémentaires (ADNc) simple brin en utilisant une amorce poly(T) et une polymérase transcriptase inverse. Les ADNc simple brin sont amplifiés en ADNc double brins qui sont ensuite séquencés en utilisant la technologie Illumina HiSeq pour générer des courtes lectures de 100 pb. Les lectures obtenues sont alignées sur le génome de référence et leur nombre est comptabilisé pour chaque gène. Pour prendre en compte les biais existants dans les données de séquençage dus aux différences de profondeur de séquençage et des longueurs des gènes, il convient d'utiliser une unité normalisée. Le choix du type de normalisation est déterminant pour pouvoir comparer les données entre les différents gènes, sources de données, espèces et, dans le cas présent, stades de développement (S., Zhao *et al.*, 2020). Au cours de ce travail c'est le TPM, *transcripts per million*, qui a été utilisé. Calculé selon la formule indiquée ci-dessous, le TPM prend en compte la longueur des gènes et permet de comparer des échantillons différents.

$$TPM = A \times \frac{1}{\sum(A)} \times 10^6 \quad \text{avec} \quad A = \frac{Nombre\ de\ lectures\ alignées\ sur\ le\ gène \times 10^3}{longueur\ du\ gène\ en\ pb}$$

En cumulant les trois stades étudiés pour chaque espèce, 15 256, 21 142 et 52 254 gènes sont exprimés (TPM>1) dans le grain chez *Brachypodium*, le maïs et le blé hexaploïde, respectivement. Cela représente 44% du total des gènes annotés pour les trois espèces (de 28% chez le maïs à 250DD à 55% chez Brachypodium à 500DD). Les gènes les plus exprimés (top100) sont logiquement enrichis en GO lié à l'activité de développement et de remplissage du grain.

Données de méthylation

L'analyse de la méthylation a été réalisée par séquençage bisulfite pour *Brachypodium* et le maïs parallèlement à l'étude de l'expression, aux mêmes stades de développement. L'ADN génomique a été extrait, fragmenté puis converti au bisulfite. Après amplification par PCR contrôlée destinée à déterminer le nombre optimal de cycles d'amplification pour obtenir une librairie représentative de la diversité avec un minimum de lectures dupliquées, l'ADN converti a été séquencé en *paired-ends* (2×100 pb) sur une plateforme Illumina HiSeq 4000. Les lectures ont été alignées sur les génomes de référence de *Brachypodium* et de maïs en utilisant Bismarck (Krueger and Andrews, 2011) et Bowtie2 (Langmead *et al.*, 2009). Cette analyse a permis d'obtenir les valeurs de méthylation pour les trois contextes (CG, CHG et CHH). Le package R methylKit a

été utilisé pour annoter la méthylation des promoteurs (400 pb autour du TSS) et des gènes (exons + introns).

Le taux de méthylation habituellement utilisé pour estimer la méthylation d'une séquence correspond au rapport entre le nombre de cytosine méthylées (mCs) sur le nombre total de cytosines méthylées et non méthylées (Cs). Il constitue une valeur normalisée qui ne tient pas compte du nombre total de cytosine. Or, en faisant l'hypothèse que le nombre de mCs peut avoir un impact sur la régulation de l'expression d'un gène, il semblait important de tenir compte de ce facteur pour étudier le lien entre la méthylation des gènes et leur expression. Pour cela, une nouvelle approche de calcul de la valeur de la méthylation pour une région donnée, promoteur ou corps du gène, appelée rpd, *read per density* (lecture par densité), a été développée, en pondérant le taux de méthylation par le nombre total de mCs.

$$rpd = mCs \times \frac{mCs}{mCs + Cs}$$

La méthylation des contextes CG, CHG et CHH de *Brachypodium* et du maïs est analysée à deux niveaux. Une première analyse globale qui agrège l'ensemble des données relatives aux promoteurs (200 pb de chaque côté du site de début de la transcription, *Transcript Start Site*, TSS) et aux gènes, incluant les exons et les introns (*gene body methylation*), indique que les niveaux de méthylation varient à la fois selon l'espèce et le contexte (annexes : Supplementary Table 7). Le maïs présente un niveau de méthylation plus élevé pour les contextes CG et CHG, ce qui peut s'expliquer par la taille du génome et le contenu important (85% du génome) en séquences répétées dont de nombreux éléments transposables. À l'inverse, *Brachypodium* présente une méthylation plus élevée pour le contexte CHH. Une seconde analyse différencie les données des promoteurs de celles des gènes. Le nombre de gènes méthylés au niveau des promoteurs (méthylation>5%) chez *Brachypodium* est de 3 316, 2 373 et 2 402, pour les contextes CG, CHG et CHH respectivement. Les chiffres de méthylation des gènes présentent des valeurs plus élevées, à savoir 14 970, 8 987 et 9 540 pour les contextes CG, CHG et CHH respectivement. Chez le maïs, les chiffres dans les contextes CG, CHG et CHH s'élèvent à 9 872, 9 406 et 3359 pour les promoteurs et 25 501, 18 941, 3 935 pour les gènes.

Données de SNP

Les SNP ont été inférés chez *Brachypodium*, le maïs et le blé sur la base de données disponibles dans la bibliographie (Gordon *et al.*, 2017; Bukowski *et al.*, 2018; Pont, *et al.*, 2019). Le nombre de SNP par gènes est normalisé en fonction de la taille du gène et exprimé en SNP par kilobase.

4.3.2 Comparaison interspécifique de la plasticité et de la régulation des gènes

Méthode de comparaison des données

L'ensemble des données omiques (expression, méthylation, SNP) ont été comparées entre chaque compartiments génomiques définis au niveau du caryotype (Inversions), du statut des gènes (conservés et spécifiques / paires et singletons) et des sous-génomes (fractions LF et MF) (Figure 41).

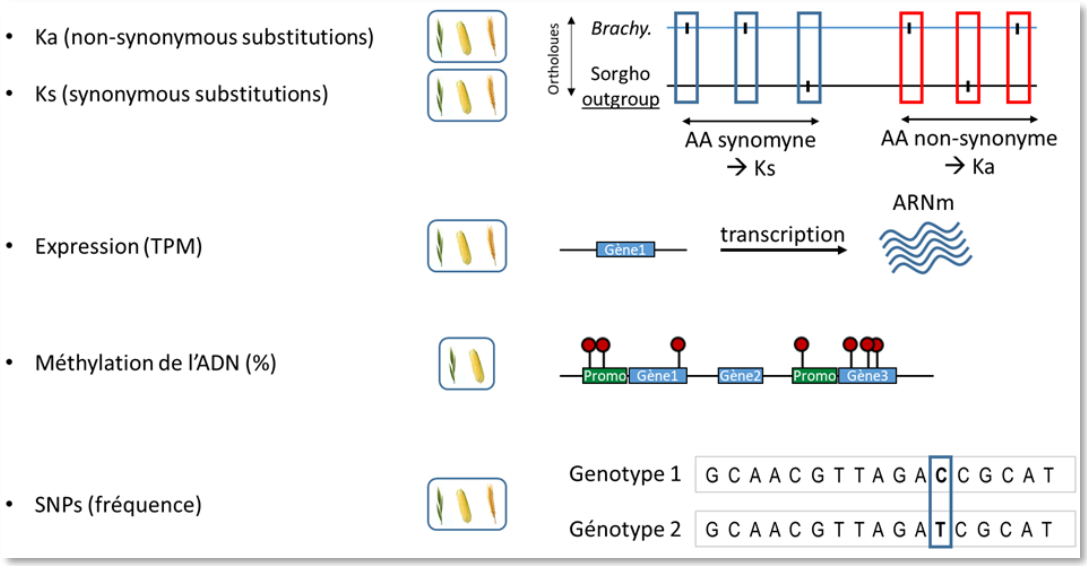


Figure 41 : Présentation synthétique des données omiques utilisées.

Les différences observées ont été validées statistiquement en comparant les deux distributions obtenues (pour chaque compartiments génomiques) par les tests de Kolmogorov-Smirnov et d'Anderson-Darling et les moyennes par les tests de Student et de Kruskal-Wallis. Si, au minimum, deux tests sur quatre indiquaient une différence entre les distributions alors cette différence était considérée comme la traduction d'un effet biologique. Les tendances observées ont été comparées entre espèces afin de définir leur degré de généricité.

L'exemple de l'étude des différences de Ks entre les gènes des compartiments LF et MF illustre la méthode mise en place (Figure 42.A). Cet exemple permet de conclure à une différence significative du taux de mutation synonyme entre les compartiments LF et MF, LF présentant un Ks significativement plus élevé que MF pour les trois espèces. Cette différence est validée par les quatre tests statistiques. Ces différences sont matérialisées par les flèches ascendantes (Ks plus élevé pour MF) et descendantes (Ks plus réduit pour LF) dans la ligne de synthèse (Figure 42.B).

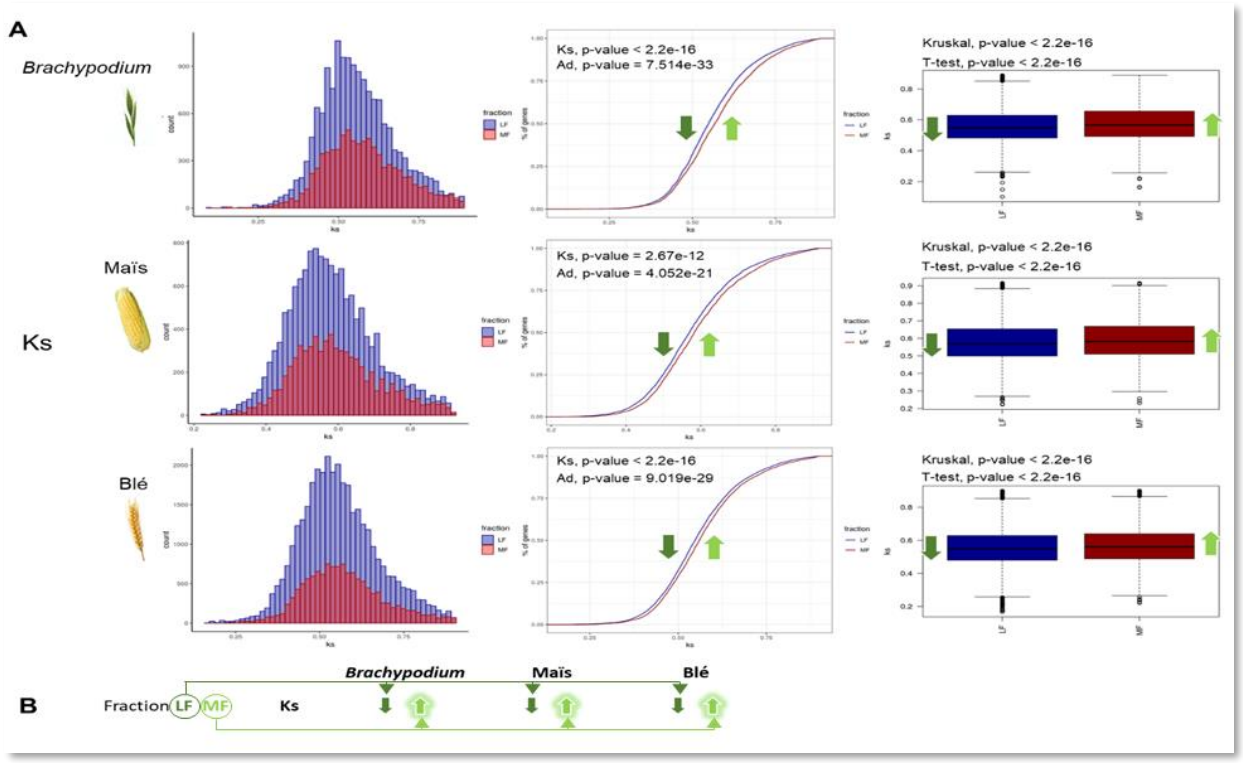


Figure 42 : Exemple d'analyse croisée entre le compartiment LF/MF et le taux de substitution synonyme des gènes appartenant aux deux compartiments. Panneau A : Pour chaque des trois espèces, Brachyopdium, le maïs et le blé, trois graphiques sont présentés. Le premier représente la distribution des gènes des deux compartiments (effectif) en fonction de leur Ks. Le deuxième graphique représente les mêmes données sous forme de pourcentage cumulé des gènes en fonction de leur valeur de Ks. L'écart entre les courbes illustre des différences dans les distributions des deux groupes ; le décalage vers la droite de la courbe représentant le Ks des gènes MF traduit une valeur de Ks supérieure pour ceux-ci par rapport aux gènes LF. Le troisième graphique représente une comparaison des valeurs médianes de Ks dans les deux ensembles. Les différences observées sont testées statistiquement en comparant les deux distributions par les tests de Kolmogorov-Smirnov, noté Ks, et d'Anderson-Darling, noté Ad, et les médianes par les tests de Student, noté T-test, et de Kruskal-Wallis, noté Kruskal. Les valeurs de p-value de ces tests sont indiquées sur les graphiques correspondants. Les flèches en vert foncé et vert clair montrent les tendances comparées entre les compartiments LF et MF. Panneau B : exemple de synthèse des résultats observés dans le panneau A qui constitue le modèle utilisé pour comparer l'ensemble des valeurs et des contextes étudiés

Résultats

Pour évaluer la généricité, j'ai comparé les résultats observés pour les trois espèces et considéré la possibilité d'un patron générique lorsque, relativement à la WGD ancestrale, lorsque les trois espèces présentent des tendances identiques, ou que deux espèces sur trois présentent des tendances identiques sans qu'il n'y ait de différences observées pour la troisième espèce (Figure 43).

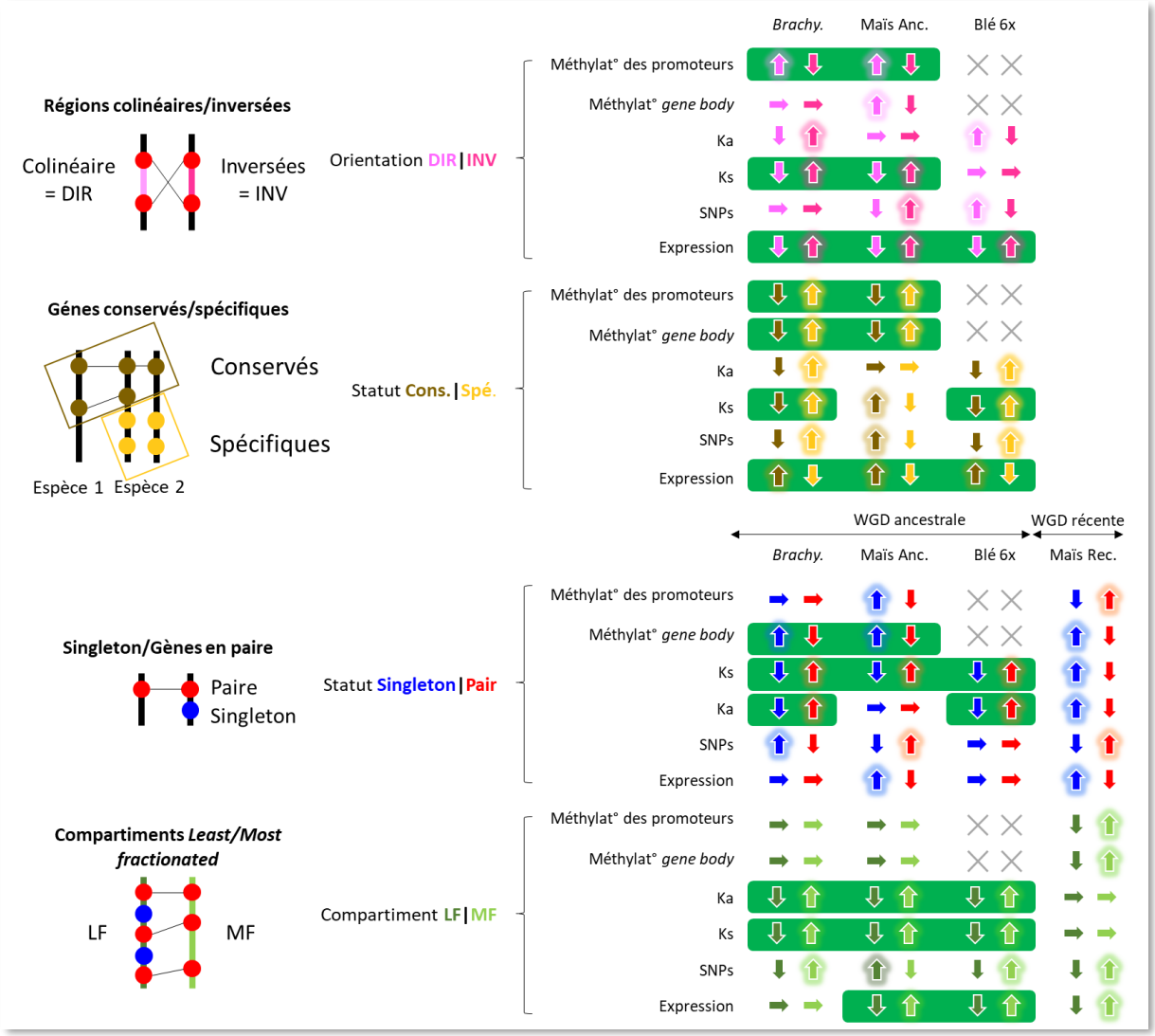


Figure 43 : Synthèse des résultats obtenus en confrontant les données omiques pour les quatre compartiments génomiques comparés. Les flèches reprennent les codes couleurs propres à chaque statut ou compartiment, elles indiquent pour chaque couple de compartiments ou statuts lesquels présentent des différences statistiquement validées (flèches vers le haut et flèches vers le bas). Les flèches horizontales indiquent qu'il n'y a pas de différence validée statistiquement. Les encadrés verts indiquent que, relativement à la WGD ancestrale, les trois espèces présentent des tendances identiques ou que deux espèces sur trois présentent des tendances identiques sans qu'il n'y ait de différences observées pour la troisième espèce.

Conséquences des inversions.

Les dynamiques des gènes présents dans les inversions sont comparées à celles des gènes des régions colinéaires (par rapport à AGK12). L'hypothèse est que les inversions peuvent impacter la régulation des gènes du fait de modifications des interactions fonctionnelles cis- et/ou trans- induits par des changements locaux dans l'ordre des gènes. En termes de mutations, les gènes présents dans les inversions présentent des valeurs Ks supérieures chez *Brachypodium* et le maïs. Au niveau des SNP les résultats sont particulièrement contrastés et les trois possibilités apparaissent : pas de différence pour *Brachypodium*, augmentation du taux chez le maïs et baisse du taux chez le blé dans les régions inversées. La tendance est, par contre, partagée entre les trois espèces pour l'expression et la méthylation des promoteurs, les gènes appartenant aux régions inversées apparaissant plus exprimés et plus méthylés que ceux des régions colinéaires.

Conséquence de l'orthologie.

La deuxième catégorie évolutive de gènes considérée est celle des gènes conservés, ou orthologues (comparativement aux gènes non conservés, ou spécifiques d'une espèce et donc retrouvés uniquement chez celle-ci et absents du reste du panel). Les gènes spécifiques sont plus méthylés, à la fois au niveau des promoteurs et des *gene body*, et moins exprimés que les gènes conservés entre espèces. En ce qui concerne les Ka, Ks et SNP il n'y a pas de dynamiques transversales entre espèces. Les gènes spécifiques de *Brachypodium* et du blé présentent des valeurs plus élevées de Ka, Ks et SNP que les gènes conservés. Ces tendances sont inverses chez les maïs hormis pour le Ks pour lequel il n'y a pas de différences significatives.

Conséquence de la paralogie.

La troisième catégorie évolutive de gènes est celle des gènes dupliqués, ou paralogues, comparativement aux singletons. Les statuts des gènes issus des deux WGD, ancestrale, et spécifique pour le maïs, sont considérés. En termes de structure les gènes issus de la WGD ancestrale conservés en paires accumulent davantage de mutations (Ka et Ks) que les singletons ce qui corrobore les résultats obtenus l'analyse du panel de 8 espèces (page 101). La situation inverse, les singletons présentant davantage de mutations que les gènes conservés en paires, est observée pour les gènes de maïs issus de la WGD datant de 5 millions d'années. Les dynamiques

sont différentes pour les SNP, les trois possibilités étant là encore visibles, une cohérence apparaissant néanmoins pour le maïs qui présente davantage de SNP dans les pairs que dans les singletons, à l'instar de la tendance observée pour les Ka et Ks des gènes dupliqués suite à la WGD ancestrale. Les niveaux d'expression ne sont pas différents, quels que soient les statuts des gènes pour Brachypodium et le blé, tandis que, chez le maïs, ils sont plus faibles chez les gènes conservés en paires pour les événements de polyploïdisation (anciens et récents). La méthylation des promoteurs ne présente pas de différence chez *Brachypodium*, à la différence du maïs qui présente une méthylation plus forte pour les singletons issus de la WGD ancestrale et plus faible chez les singletons issus de la WGD récente. La méthylation *gene body* des singletons est plus élevée que celle des gènes en paires pour les deux espèces et les deux WGD du maïs.

Conséquence de la dominance des sous-génomes.

Entre les fractions LF et MF, en termes de mutations Ka et Ks, les gènes de la fraction MF accumulent globalement davantage de mutations que ceux de la fraction LF, dans 6 observations sur 6. En ce qui concerne l'expression, les gènes de la fraction MF apparaissent plus exprimés chez le maïs et le blé, tandis que chez *Brachypodium* il n'y a pas de différence observable. Aucun biais de méthylation, ni pour les promoteurs, ni pour le corps des gènes, n'a été observé entre les fractions LF et MF ancienne, tandis que la fraction MF récente du maïs présente des niveaux de méthylation, promoteurs et *gene body*, plus élevés que ceux de la LF récente.

Résultats détaillés en annexes Supplementary Figure 8b.

En conclusion, sur la base de l'observation des effets sur la régulation observés pour chaque catégorie structurale de gènes (Figure 43), les tendances majoritaires qui apparaissent comme conservées entre les espèces étudiées sont :

- **L'ordre des gènes** (inversions vs colinéarité) : **Les gènes des régions inversées sont majoritairement moins méthylés (promoteurs) et plus exprimés** que les gènes des régions conservant l'orientation ancestrale. Pas de tendances conservées pour la méthylation *gene body*, le Ks et le Ka, et les SNP ;

- **La conservation des gènes** (gènes conservés vs spécifiques) : **Les gènes conservés sont majoritairement moins méthylés (promoteurs et *gene body*) et plus exprimés** que les gènes spécifiques. Pas de tendances conservées pour le Ks et le Ka, et les SNP ;

- **La duplication des gènes** (gènes dupliqués vs singletons) : **Les paires issues de la duplication ancestrale sont moins méthylées pour leurs *gene body*** (même tendance pour la WGD récente du maïs) et présentent des valeurs de Ks supérieures (tendance inverse pour la WGD récente du maïs) que les singletons. Pas de tendances conservées pour la méthylation des promoteurs, le Ka, l'expression et les SNP.

- **La diploïdisation des gènes (compartiment LF vs MF)** : Pour les compartiments issus de la duplication ancestrale ρ, **les valeurs de Ks et Ka des gènes MF sont supérieures** à celles des gènes LF (pas de différence pour la WGD récente du maïs). Pas de tendances conservées pour la méthylation des promoteurs et *gene body*, et les SNP. **Les gènes MF sont majoritairement plus exprimés que les gènes LF chez le maïs (pour les deux WGD) et le blé,** tandis qu'il n'y a pas de différence significative chez *Brachypodium.*

4.3.3 Cas du blé hexaploïde : plasticité et régulation des gènes des sous-génomes

Pour mieux comprendre l'effet des polyploïdisations récentes sur la régulation des gènes, j'ai considéré le modèle du blé hexaploïde. Son histoire évolutive est bien documentée et se caractérise par l'hybridation de trois sous-génomes ayant divergé à partir d'un ancêtre commun dont ils ont conservé la structure, il y a 6,5 millions d'années (Figure 44).

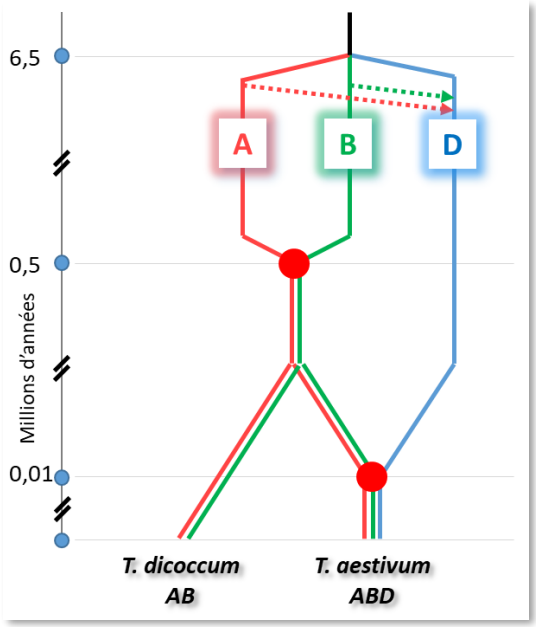


Figure 44 : Mise en place de l'hexaploïdie chez le blé. Les WGD sont représentées par des points rouges. Les sous-génomes A, B et D sont issus d'un évènement de spéciation il y a 6,5 millions d'années. Les blés diploïdes évoluent durant 6 millions d'années au cours desquelles interviennent des introgressions des génomes A et B dans le génome D. Il y a 0,5 million d'années les génomes des blés A et B s'hybrident formant le blé tétraploïde, puis il y a environ 10 000 ans, le génome D s'hybride au génome AB formant le blé hexaploïde. Adapté de El Baidouri et al., 2017.

En appliquant les mêmes méthodes que précédemment, j'ai comparé les valeurs d'expression, de Ka et de Ks, et de fréquence des SNP entre les sous-génomes A, B et D du blé hexaploïde.

Une première comparaison en prenant l'ensemble des gènes de chaque sous-génome sans a priori (Figure 45) montre que le sous-génome D présente un niveau d'expression plus élevé que celui du sous-génome B, qui lui-même est plus exprimé que le sous-génome A. A l'inverse, le sous-génome D présente des taux de SNP et des valeurs de Ks plus faibles que ceux des sous-génomes A et B, pour ces deux valeurs, respectivement B est supérieur à A et A et B ne présentent pas de différences significatives. Aucune différence significative n'est détectée pour les valeurs de Ka, ce qui laisse penser que la pression de sélection liée à la conservation des fonctions de gènes s'exerce uniformément sur les 3 sous-génomes. Globalement, le sous-génome D qui s'est hybridé il y a environ 10 000 ans au blé tétraploïde AB, présente des patrons d'expression et de mutations substantiellement différents de ceux des sous-génomes A et B qui sont en interactions au sein du blé tétraploïde depuis un demi-million d'années.

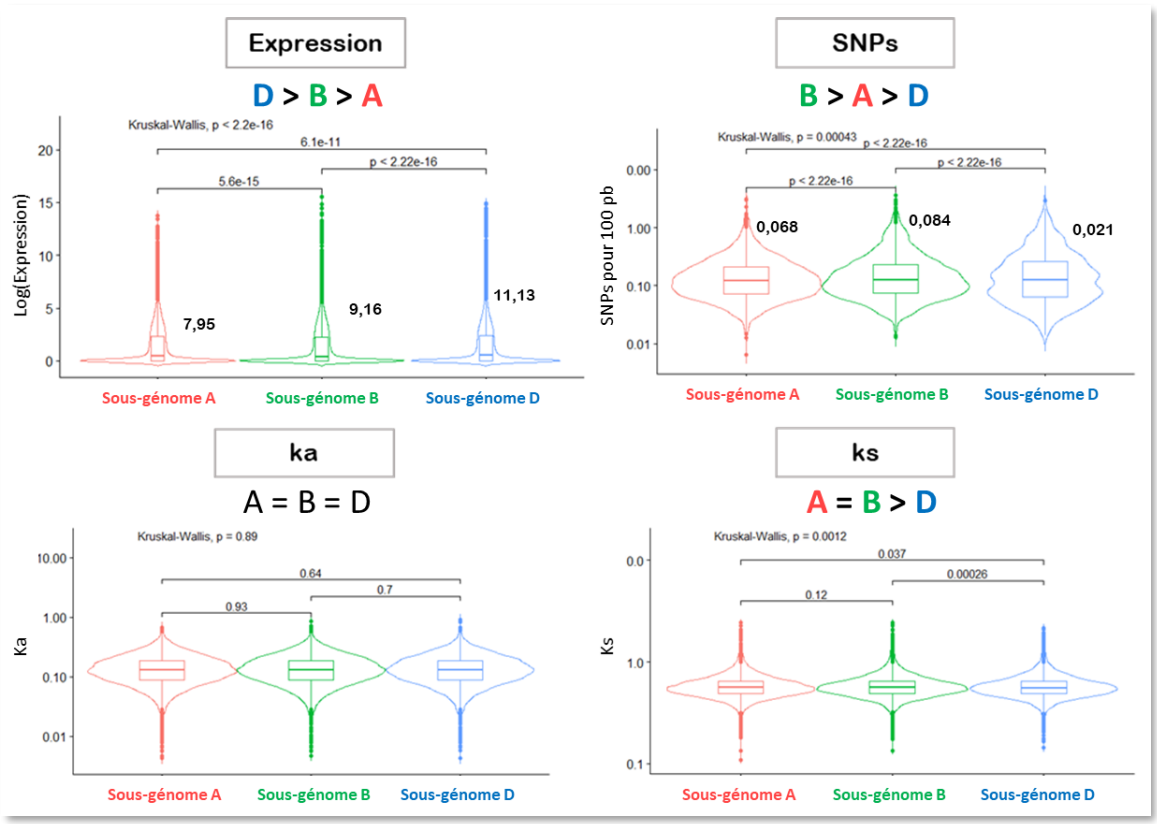


Figure 45 : Comparaison des valeurs d'expression (exprimé en log2(exp+1)), taux de SNP (exprimé en nombre de SNP pour 100 pb dans les CDS), Ka et Ks entre les sous-génomes A, B et D du blé hexaploïde. Les valeurs de p-value indiquées au-dessus des crochets traduisent des différences statistiques lorsqu'elles sont inférieures à 0,05. Les relations entre les sous-génomes sont résumées sous le nom de chaque variable.

Ensuite, l'impact des compartiments LF et MF des gènes issus de la duplication ancestrale ont été évalués. En considérant les trois sous-génomes, 6 compartiments génomiques chez le blé ont ainsi été définis : MF-A, MF-B, MF-D, LF-A, LF-B et LF-D (Figure 46). Entre les compartiments LF et MF, des différences au sein des sous-génomes A-B-D sont observées pour l'expression, le taux de SNP et le Ks tandis qu'aucun des 3 sous-génomes ne présente de différences pour le Ka. Dans le détail, les compartiments LF et MF présentent des différences pour le Ks pour les trois sous génomes A-B-D. Pour l'expression, les sous-génomes B et D présentent des différences entre compartiments LF et MF. Pour les SNP, seul le sous-génome B présente une différence entre les compartiments LF et MF. Lorsque des différences existent, c'est le systématiquement compartiment MF qui présente les valeurs plus élevées, comme observé pour *Brachypodium* et le maïs.

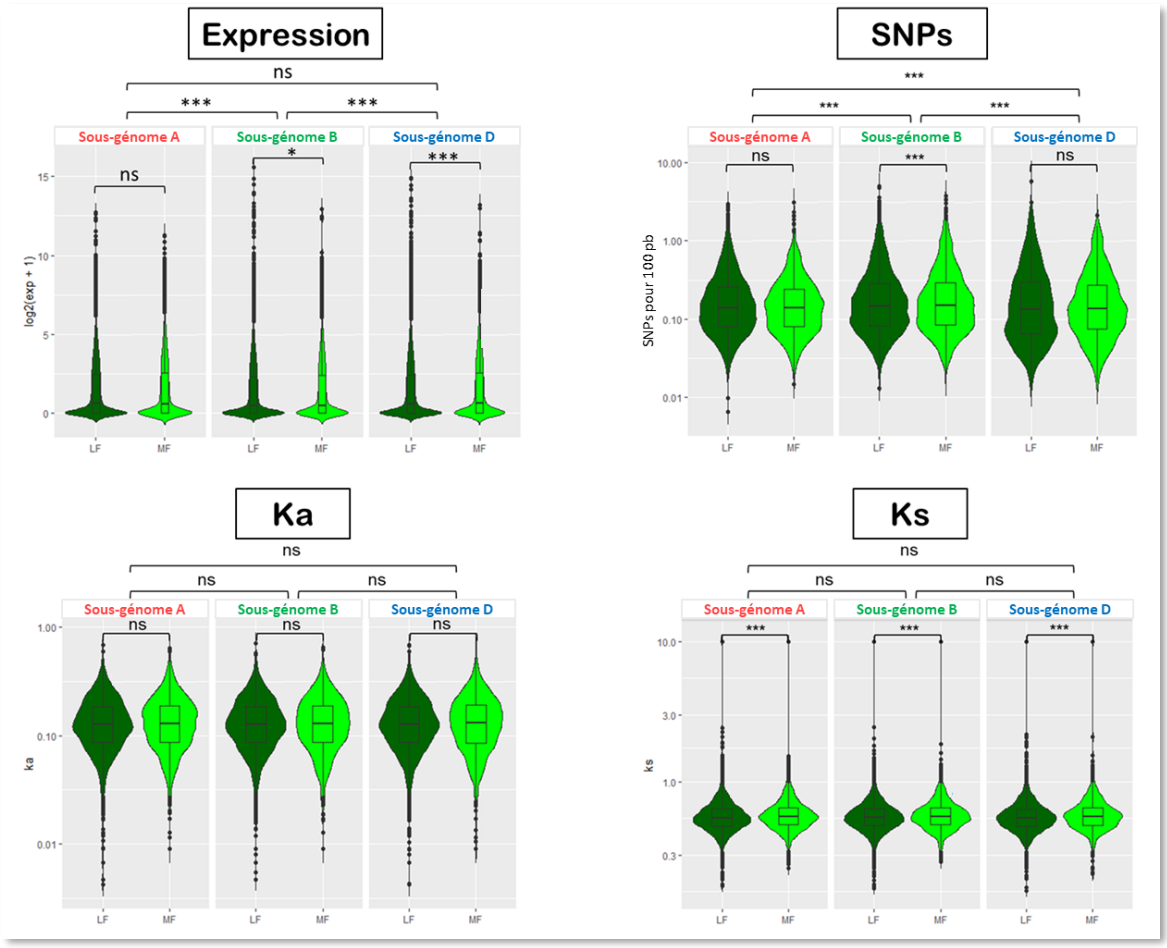


Figure 46 : Comparaison des valeurs d'Expression, taux de SNP, Ka et Ks entre les fractions LF et MF sous-génomes A, B et D du blé hexaploïde. Les valeurs au-dessus des crochets indiquent si les valeurs comparées présentent des différences statistiquement significatives. Les étoiles indiquent que les différences sont significatives et leur nombre

traduit une puissance croissante de la valeur du résultat du test statistique. ns indique le test n'est pas significatif. Les crochets au-dessus des boîtes correspondent aux comparaisons entre les trois sous-génomes pour une fraction LF ou MF donnée. Les crochets dans les boîtes correspondent aux comparaisons entre les fractions LF ou MF au sein d'un sous-génome donné.

Les comparaisons deux à deux entre les compartiments LF et MF des trois sous-génomes A, B et D ne montrent pas de biais pour les substitutions synonymes, ni pour les substitutions non synonymes. Par contre, des différences existent pour les taux de SNP et pour les niveaux d'expression. Les gènes LF du sous-génome D (LF-D) accumulent moins de SNP que les gènes LF des sous-génomes A et B (LF-B > LF-A > LF-D). Le même schéma a été observé pour les gènes MF où MF-B > MF-A > MF-D. Ces résultats sont similaires à l'étude sans a priori des gènes des sous-génomes. L'analyse des données d'expression présente un profil plus complexe. Alors que le sous-génome B est le moins exprimé dans les compartiments LF et MF, le sous-génome A apparaît comme le plus exprimé dans LF tandis que les sous-génomes A et D sont les plus exprimés dans les compartiments MF (annexes : Supplementary Table 8).

4.3.4 Plasticité et régulation des gènes retenus au cours de l'évolution

4.3.4.1 Plasticité et régulation des gènes conservés

Pour analyser plus finement les impacts de l'histoire évolutive et des structures des génomes sur la régulation des gènes (expression et méthylation), l'analyse s'est focalisée sur les gènes parfaitement conservés au cours de l'évolution, c'est-à-dire (i) présents chez les trois espèces, et (ii) retenus en paires chez le maïs et en triplets chez le blé. Des modules d'expression et de méthylation pour chaque gène ont été définis en concaténant leur statut d'expression/méthylation (0 pour non-exprimé/non-méthylé, 1 pour exprimé/méthylé) à chacun des trois stades de développement. A titre d'exemple, le module 0-1-0 correspond à une expression ou méthylation uniquement au stade 2 (250DD chez le blé). Un répertoire de 1997 gènes a été défini ; ces gènes présentant la relation génique 1-2-3 gènes chez *Brachypodium*, le maïs et le blé, respectivement, correspondant aux trois niveaux de ploïdie : diploïde, tétraploïde et hexaploïde. J'ai comparé les modules d'expression de ces gènes pour les trois espèces (Figure 47).

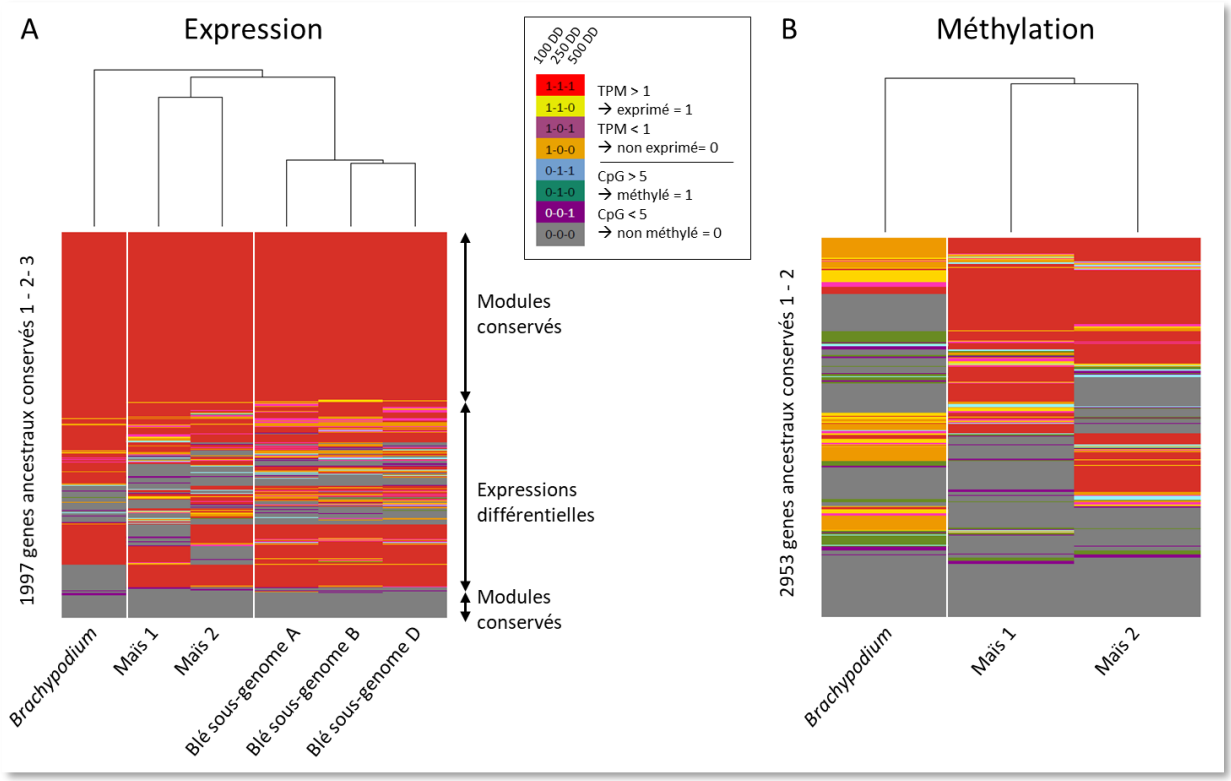


Figure 47 : Analyse du scénario de conservation ou de divergence des modules d'expression (A) et de méthylation (B) entre les gènes orthologues de Brachypodium (1 copie) , du maïs (2 copies) et du blé hexaploïde (3 copies). Les patrons de conservation/divergence sont observés à l'échelle interspécifique et à l'échelle intraspécifique des sous génomes, compartiments issus de la tétraploïdisation du maïs (Maïs 1, Maïs 2), et sous-génomes A, B et D du blé hexaploïde.

Ces résultats indiquent que près de la moitié des gènes ancestraux conservent le même profil d'expression dans les trois espèces (43% exprimés dans toutes les espèces et 6% non exprimés dans les trois espèces). Chez le maïs, 70 % des paires partagent le même module d'expression (58 % 1-1-1 et 12 % 0-0-0), tandis que pour 13 % des paires lorsque l'une des copies est exprimée, l'autre ne l'est pas (profil On/Off). Des schémas similaires sont observés parmi les triplets d'homéologues du blé, 78 % des triplets partageant le même module d'expression, dont 64 % correspondant aux modules 1-1-1 et 14 % aux modules 0-0-0. Seuls 5% des homéologues du blé présentent un profil d'expression On/Off entre les trois sous-génomes.

La même stratégie a été appliquée aux données de méthylation en comparant *Brachypodium* et le maïs, à savoir identification d'un répertoire de 2953 gènes présentant une relation parfaite 1 pour 2, entre *Brachypodium* et le maïs. Ces gènes ont été utilisé pour évaluer les différences de

méthylation de l'ADN entre les gènes conservés (seuil minimum de méthylation = 5 CpG à 10X de couverture). L'analyse du méthylome montre des différences plus marquées entre les trois stades de développement, en particulier chez *Brachypodium* où 46 % des gènes étudiés présentent un profil distinct entre les stades, probablement en raison des changements de méthylation de l'ADN particulièrement intenses au moment des premiers stades embryonnaires du développement du grain. Chez le maïs, 34% des gènes étudiés montrent des différences de méthylation entre les trois stades de développement. La comparaison entre les ohnologues de *Brachypodium* et de maïs montre que le niveau global de méthylation des gènes est plus élevé chez le maïs (22% avec 1-1-1 et 23% avec 0-0-0 entre les duplicatas de maïs) que chez *Brachypodium* (5% avec 1-1-1 et 49% avec le module 0-0-0). Seulement 16% des gènes ancestraux conservent les mêmes patrons de méthylation entre les deux espèces, dont 14% correspondant au module 0-0-0 (gènes non méthylés dans les trois stades). Si l'on compare les paires de gènes du maïs, ce pourcentage augmente à 45% dont 22% des paires de gènes présentant le module 1-1-1 et 23% le module 0-0-0.

4.3.4.2 Plasticité et régulation des paires de gènes

En complément de l'étude de la régulation des gènes conservés, cette section présente spécifiquement les paires de gènes homéologues issus, soit de la WGD ancestrale ρ, pour les trois espèces, soit de la duplication spécifique du maïs il y a 5 millions d'années. Au sein des paires, les gènes sont considérés en fonction de leur appartenance à l'un ou l'autre compartiment LF ou MF. Les niveaux d'expression et de méthylation des paires sont comparés, pour identifier les gènes différentiellement exprimés, ou DEG (*Differentially Expressed Genes*), et gènes différentiellement méthylés, ou DMG (*Differentially Methylated Genes*). Dans la littérature, la plupart des études indiquent que le patron d'expression majoritaire pour les gènes en paires est qu'une copie est exprimée et l'autre ne l'est pas. Ici, les résultats apparaissent plus nuancés (Figure 48). Seulement 43% des gènes observés (1100/2575) issus de la WGD ancestrale, en cumulant les paires de gènes de *Brachypodium* du maïs et du blé, et 52% des gènes observés (1161/2254) issus de la WGD récente du maïs, sont différentiellement exprimés. Parmi les gènes différentiellement exprimés, les gènes surexprimés appartiennent majoritairement à la fraction MF (61% MF, 39% LF) lorsqu'ils

sont issus de la WGD ancestrale. Lorsque l'on considère les gènes du maïs, issus de la WGD récente, différentiellement exprimés, cette différence est plus ténue (46% MF, 54% LF).

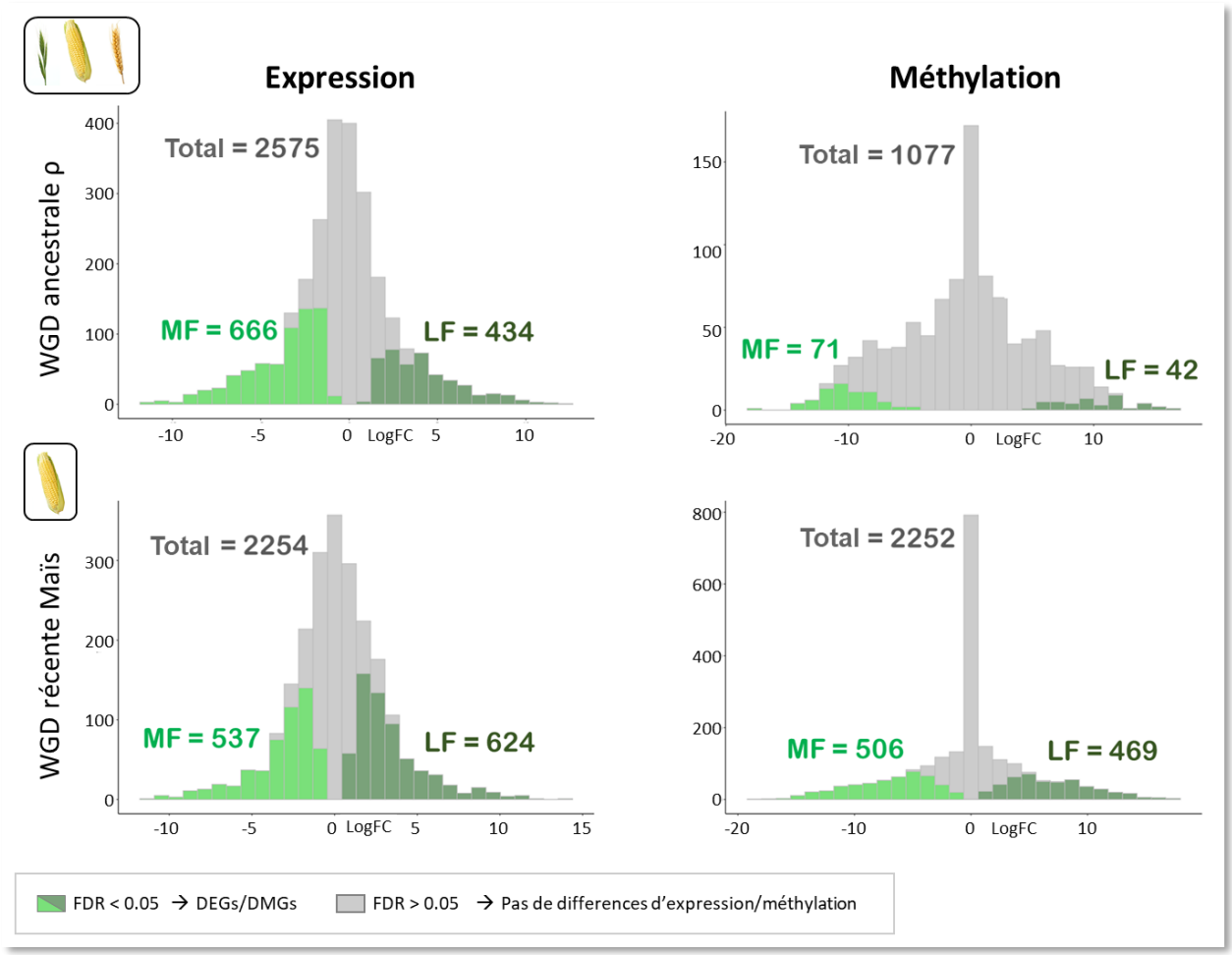


Figure 48 : comparaison des niveaux d'expression et de méthylation des copies homoéologues issues des WGD ancestrale (partie supérieure) et spécifique du maïs (partie inférieure). Les différences d'expression et de méthylation sont exprimées en log du Fold Change. Ici, le fold-change est le rapport au sein d'une paire de gènes homéologues entre les niveaux d'expressions des deux copies. Il est exprimé en log (logarithme en base 2) afin de rendre symétriques les rapports par rapport à 0. Par exemple, la copie A gène ayant un fold-change de 1 (respectivement -1) par rapport à la copie B signifie que la copie A est deux fois plus (respectivement moins) exprimée que la copie B.

L'analyse du méthylome des gènes dupliqués montre également des résultats sensiblement différents entre les deux évènements de duplication considérés, l'ancien et le récent. Seules 10% des paires de gènes issues de la duplication ancestrale présentent des niveaux de méthylation différents (113/1077). Ce chiffre faible s'explique probablement par l'intervalle de temps très long (90 millions d'années) durant lequel de multiples évènements de méthylation

ont pu se surimprimer à la méthylation post-WGD, rendant difficile la détection d'une signature propre à l'évènement initial. Par contre, dans le cas de la duplication récente du maïs (5 millions d'années), 43% des gènes (975/2252) présentent des différences de méthylation de l'ADN. Il est notable que les copies hyperméthylées appartiennent dans le cas présent autant au compartiment LF qu'au compartiment MF, respectivement 48% et 52%.

4.3.5 Analyse multiomiques

En complément de l'étude de l'analyse des données omiques en fonction des différentes catégories de gènes identifiées (contenus dans les inversions, conservés/spécifiques, LF/MF, singletons/paires), les relations entre les différentes données omiques elles-mêmes ont été étudiées. Les analyses ont notamment porté sur les relations entre la dynamique des mutations, estimée par le Ka, et la méthylation des gènes, d'une part, et les liens entre l'expression des gènes et l'ensemble des autres données omiques (Ka, méthylation des promoteurs et méthylation gene body notamment) d'autre part.

Pour cela les données sont organisées en quantiles. Les données (Ka, méthylation des promoteurs et méthylation gene body) sont ordonnées selon une distribution croissante et les demi-déciles partagent cette distribution en 20 parties. Ainsi, pour la méthylation par exemple le premier demi-décile inclut les 5% des gènes du jeu de données présentant les valeurs de taux de méthylation les plus faibles. Les valeurs de méthylation (i) et d'expression (ii) moyennes sont calculés pour les gènes appartenant à chaque demi-décile (Figure 49). Les tendances présentées ci-dessous sont celles du le maïs. Elles sont semblables pour *Brachypodium*.

L'analyse a permis de mettre en évidence une corrélation négative entre le Ka et la méthylation des gènes (Figure 49.A). Cette information accrédite l'hypothèse que plus un gène est méthylé, moins il sera affecté par les mutations.

Une corrélation inverse a été mise en évidence entre le taux de substitutions non synonymes Ka et l'expression des gènes (Figure 49.B). Cela indique que les gènes présentant des taux de substitution non synonymes élevés sont moins exprimés et, inversement, que les gènes les plus exprimés présentent des valeurs de Ka plus faibles. Autrement formulé, les gènes les plus conservés en termes de séquence sont les plus exprimés.

Aucun lien clair n'a été établi entre l'expression des gènes et les valeurs Ks ou les densités en SNP.

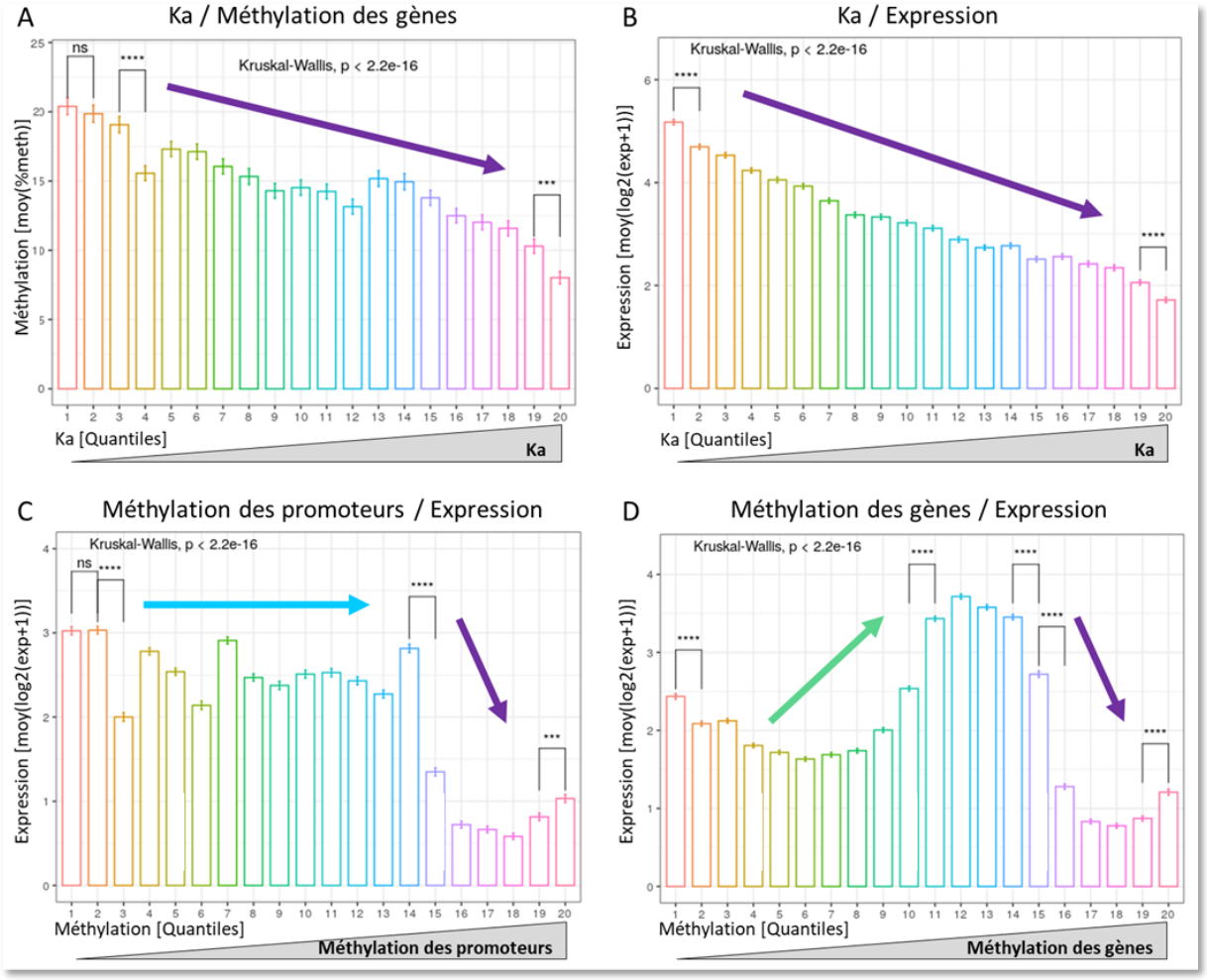


Figure 49 : Relations entre données omiques chez le maïs. Les valeurs moyennes sont calculées pour chaque demi-décile. Les demi-déciles sont classés dans l'ordre croissant de la valeur considérée. A : Méthylation des gènes en fonction du Ka. B : Expression en fonction du Ka. C : Expression en fonction de la méthylation des promoteurs. D : Expression en fonction de la méthylation des gènes (gene body). Flèche verticale : absence de corrélation / Flèche ascendante : corrélation positive / Flèche descendante : corrélation négative.

En ce qui concerne les relations entre méthylation et expression, deux tendances différentes sont apparues, selon que les promoteurs ou les gènes eux-mêmes soient considérés. Une méthylation des promoteurs faible à modérée correspondant à 70% des gènes considérés (jusqu'à Q14, demi-décile correspondant à un taux de méthylation de 53%) n'affecte par le niveau d'expression, alors que pour les 30% de gènes présentant les plus forts taux de méthylation des promoteurs (taux de méthylation supérieur à 53%), l'expression chute brutalement (Figure 49.C). L'effet de seuil est très marqué. Concernant la méthylation des gènes (*gene body*), la tendance est différente. Pour

les 60% (Q12) des gènes les moins méthylés (taux de méthylation inférieur à 30%), la tendance générale est une corrélation positive entre la méthylation et le niveau d'expression. En revanche, elle est négativement corrélée pour les gènes hautement méthylés, Q13 et plus, avec à nouveau un effet de seuil à partir de Q16, correspondant au quart des gènes les plus méthylés qui présentent des taux de méthylation de supérieurs à 58% (Figure 49.D).

En complément de cette approche comparant des valeurs moyennes de méthylation et d'expression de population de gènes, une approche multivariée mise en œuvre dans mixOmics a permis de réaliser une analyse « gène à gène ».

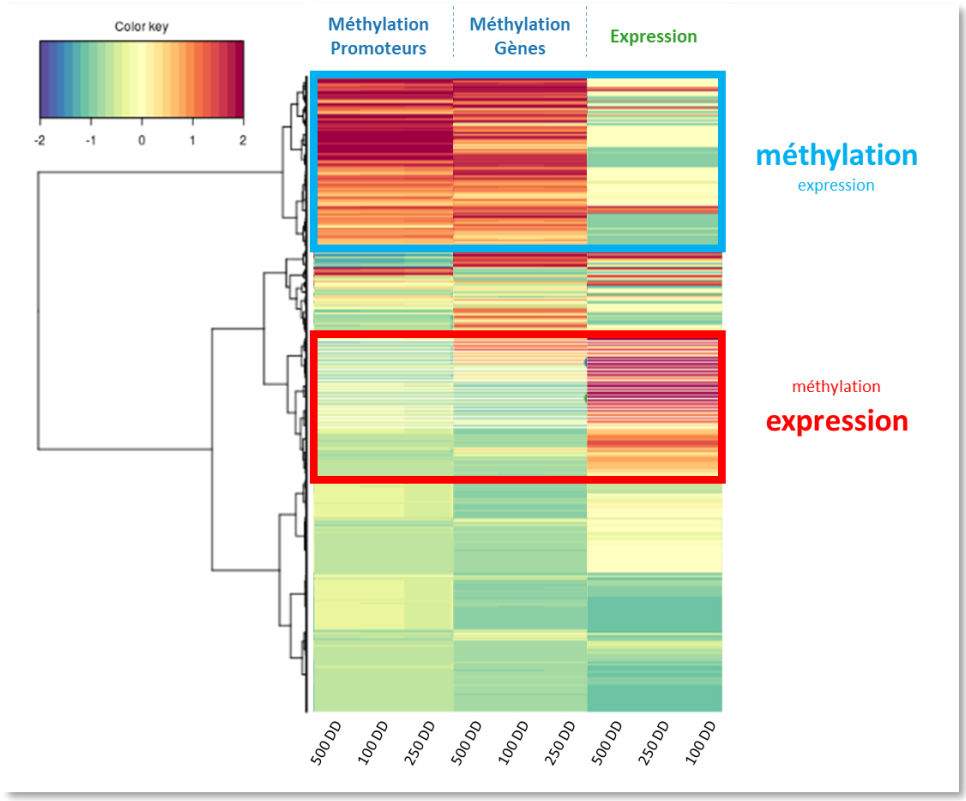


Figure 50 : Sous-ensemble de l'intégration des données d'expression et de méthylation des promoteurs et des gènes dans les trois contextes, pour les trois stades de développement, chez le maïs. L'intégration des données est faite par réduction de dimension utilisant Mixomics. Ce sous-ensemble représente les données de méthylation CpG des promoteurs et des gènes, et les niveaux d'expression associés. Chaque ligne correspond à un gène. En fonction de l'ensemble des données considérées, l'intégration des données permet de grouper les gènes en fonction des proximités entre les données omiques qui leur sont associées (expression et méthylation dans les 3 stades du développement du grain). Ce regroupement est modélisé par le dendrogramme de gauche. Les encadrés bleu et rouge indiquent les groupes de gènes, respectivement, hyperméthylés et faiblement exprimés, et hypométhylés et fortement exprimés. Les résultats complets de cette analyse ainsi que celle de Brachypodium se trouvent en annexes.

Dans l'ensemble, cette approche regroupe les échantillons selon les stades de développement et les variables omiques (expression, contexte et type de méthylation), à l'exception de la méthylation CG des promoteurs à 500DD et de la méthylation CHH des gènes à 250DD chez *Brachypodium*. Dans le cadre de ce travail qui vise à explorer les liens entre données omiques, cette analyse permet d'identifier des gènes présentant deux profils particuliers : hypométhylés et fortement exprimés, hyperméthylés et faiblement exprimés (Figure 50). Cela confirme la tendance mise en évidence précédemment.

Pour mieux caractériser ces gènes aux profils particuliers, l'expression en fonction de la méthylation (exprimée en rpd, read per density) est représentée. Le résultat confirme que, bien qu'il n'y ait pas de corrélation marquée à l'échelle du génome entier, il y a pour certains gènes, dits « extrêmes » ou « outliers », un antagonisme entre niveau d'expression et niveau de méthylation de l'ADN. Ainsi, les gènes fortement exprimés ne sont pas ou sont peu méthylés au niveau des promoteurs, tandis que les gènes fortement méthylés ne sont en majorité pas exprimés ou le sont faiblement (Figure 51). Cette conclusion vient supporter l'hypothèse que la méthylation joue un rôle clé dans la répression de l'expression des gènes.

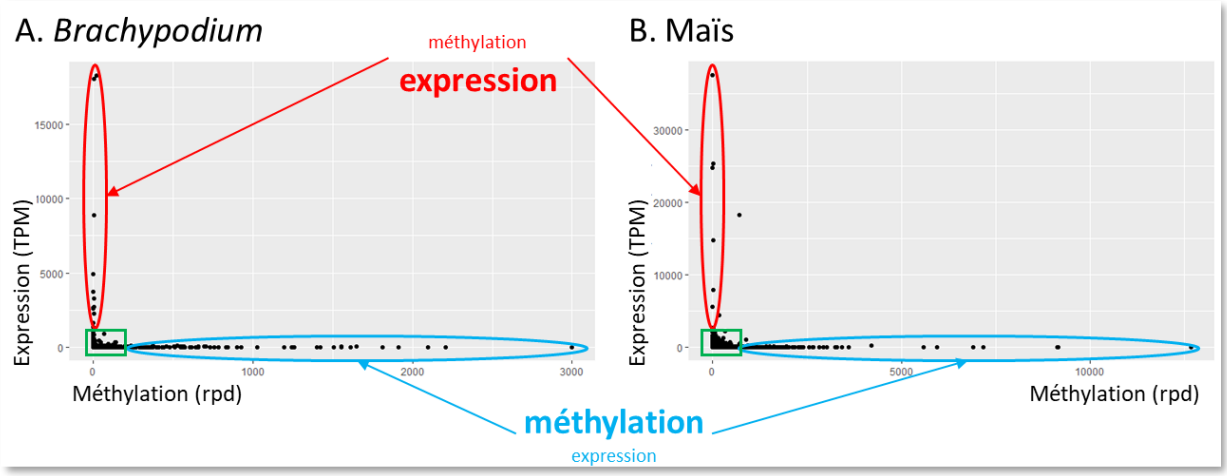


Figure 51 : Expression des gènes en fonction de leur niveau de méthylation (rpd). Les profils de Brachypodium et du maïs sont similaires. La majorité des gènes voient leurs valeurs groupées à proximité de l'origine et ne présentent pas de corrélation particulière entre méthylation et expression. Cependant, une fraction des gènes hyperméthylés et faiblement exprimés, ou hypométhylés et fortement exprimés sont identifiés.

L'analyse des GO de ces gènes *outliers* montre que, dans l'ensemble chez *Brachypodium*, les gènes hypométhylés et surexprimés correspondent à des fonctions de signalisation cellulaire

(transporteurs transmembranaires, ligase) (Annexes *Supplementary Figure* 2d). Les gènes hyperméthylés et sous-exprimés sont impliqués dans des fonctions liées à la liaison des purines/ribonucléotides, à l'activité hydrolase, à la liaison des ions, à l'activité des transporteurs, à la liaison des protéines de choc thermique (Annexes *Supplementary Figure* 2e). Il est notable que les gènes des protéines de choc thermique sont contrôlés par la méthylation de l'ADN. La dérégulation de la méthylation de l'ADN activée par le stress thermique serait donc susceptible de réactiver ces gènes.

4.3.6 Conclusions

- **L'étude de l'impact de la polyploïdie sur la régulation des gènes, dans nos conditions expérimentales au cours du développement du grain, indique que sur la base des différents attributs des gènes, à savoir, conservés/spécifiques et singletons/gènes en paires, et des compartiments inversés/colinéaires et LF/MF :**

  - Les gènes des régions inversées sont majoritairement moins méthylés (promoteurs) et plus exprimés que les gènes des régions conservant l'orientation ancestrale.

  - Les gènes conservés sont majoritairement moins méthylés (promoteurs et gene body) et plus exprimés que les gènes spécifiques.

  - Les singletons présentent majoritairement moins de mutations et sont plus méthylés que les gènes en paires.

  - Les gènes LF présentent majoritairement moins de mutations et sont moins exprimés que les gènes MF. Il n'y a pas de biais de méthylation entre les gènes des deux compartiments.

- **L'analyse spécifique des paires de gènes montre que seulement 50% des paires présentent des différences d'expression et de méthylation.**
- **Au niveau interspécifique, il apparaît que 50% des gènes ohnologues ont conservé le même patron d'expression (blé, maïs, *Brachypodium*). Ils présentent par contre des patrons de méthylation variables (maïs, *Brachypodium*).**
- **L'analyse conjointe des données omiques montre que**

  - Les gènes qui présentent le moins de mutations (Ka) sont les plus méthylés et les plus exprimés

  - Il existe une corrélation négative entre la méthylation de l'ADN et l'expression des gènes pour un sous-ensemble de gènes présentant des niveaux de méthylation et d'expression extrêmes. Cette conclusion soutient l'hypothèse que la méthylation de l'ADN joue un rôle clé dans le contrôle de l'expression des gènes.

4.4 Revue des conclusions au regard de la bibliographie

De nombreuses publications ont décrit les évènements de polyploïdie qui ont façonné les différentes espèces végétales. Elles ont mis en évidence des dynamiques multiples qui affectent les génomes aux différents niveaux de structure et de régulation. Selon les travaux et les espèces de plantes considérées, les phénomènes décrits sont parfois divergents, voire opposés, tels que le remodelage intense des chromosomes ou le maintien des structures homéologues, la dominance structurale d'un sous-génome ou, au contraire, des pertes de gènes équilibrées, la dominance en termes d'expression ou la mise en jeu de l'expression des gènes de l'un ou l'autre sous-génome en fonction des tissus ou de leurs stades de développement. Au-delà du travail de description, des hypothèses ont été avancées pour expliquer ces dynamiques, en prenant notamment en compte l'âge de l'évènement de polyploïdisation et le degré de similarité des génomes des progéniteurs à l'origine de l'événement de polyploïdie.

Le temps écoulé depuis la WGD est un facteur crucial à considérer pour comprendre le cheminement évolutif post-polyploïdie. L'étude d'un évènement récent, moins d'un million d'années, ou a fortiori d'un polyploïde synthétique, sur quelques générations, permettra d'examiner les mécanismes immédiatement mis en place post-polyploïdie, ou leur absence le cas échéant, mais elle devra être considérée comme l'image d'un état transitoire. Un événement plus ancien, entre 1 et 15 millions d'années, permet d'observer un processus plus avancé au cours duquel plusieurs mécanismes se sont combinés. Il est raisonnable de considérer cet état comme relativement stable, quoique restant non achevé tant que la plante n'aura pas retrouvé une structure diploïde (diploïdisation). Un évènement plus ancien, plus de 15 millions d'années, permet d'observer le processus complet jusqu'à la diploïdisation, mais d'autres phénomènes récents, tels que des duplications de gènes en tandem ou des insertions d'éléments transposables, la domestication ou la sélection au sens large, peuvent masquer les signatures propres à l'évolution post-polyploïdie. Il est à noter que certaines espèces modernes combinent plusieurs événements de polyploïdisation, parfois désignés par les termes de duplication récente et de paléoduplication, dont les mécanismes peuvent être étudiés en parallèle et comparés.

Le degré de similarité des génomes qui s'hybrident pour former un génome polyploïde est fréquemment avancé comme un élément déterminant l'évolution des deux sous-génomes dans le néopolyploïde formé. A cet égard, les situations sont multiples, depuis la duplication à l'identique du génome dans le cas d'une autopolyploïdie stricte, jusqu'à l'hybridation d'espèces ayant évolué séparément pendant plusieurs millions d'années dans le contexte de l'allopolyploïdisation. Dans le premier cas, les évolutions post-polyploïdie seraient permises par le relâchement de la sélection purificatrice sur l'une des copies des paires de gènes issues de la WGD (Cheng *et al.*, 2018), et par l'insertion d'ET impactant les niveaux de méthylation globaux (Zhang *et al.*, 2015) ou certains gènes directement, notamment des gènes impliqués dans des processus adaptatifs (Baduel *et al.*, 2019). Toutefois, les modifications de la structure et de la régulation des deux sous-génomes à partir d'un génome unique ayant formé l'autopolyploïde demeurent encore énigmatiques. Dans le cas de l'allopolyploïdie, les génomes des progéniteurs peuvent s'être différenciés par des mécanismes relativement rapides tels que la méthylation, l'insertion d'éléments transposables, des duplications en tandem ou au contraire des pertes de gènes. Autant d'évènements susceptibles de modifier la régulation de voies métaboliques. Toutes ces différences entre les progéniteurs sont considérées comme des explications possibles des différences de structure ou de régulation observées entre les deux sous-génomes au sein de l'allopolyploïde formé.

Les recherches visant à comprendre les mécanismes post-polyploïdie ont exploré les phénomènes de perte de gènes et, son corollaire, la conservation de gènes en paires, les dynamiques d'accumulation des mutations de la séquence, les différences d'expression, de méthylation, de contenus en éléments transposables entre les sous-génomes, et les différences d'expression entre copies au sein des paires conservées au cours de l'évolution. Les dynamiques d'évolution des gènes, selon qu'ils soient conservés en paires ou retournent à l'état de singleton, ou selon qu'ils soient des gènes conservés ancestralement ou spécifiques d'une espèce, ont été moins étudiées. De même, les études appliquant une même méthodologie à plusieurs espèces sont moins fréquentes dans la bibliographie que celles portant sur une seule espèce.

Je propose ici de confronter les résultats obtenus dans le cadre de ce travail sur les céréales aux résultats obtenus sur diverses espèces, dont *Arabidopsis*, le coton, le maïs, le soja, *Brassica rapa*

et *Brassica oleracea* progéniteurs paléopolyploïdes de *Brassica napus*, qui présentent des trajectoires évolutives différentes en réponse aux évènements de polyploïdie (Figure 52).

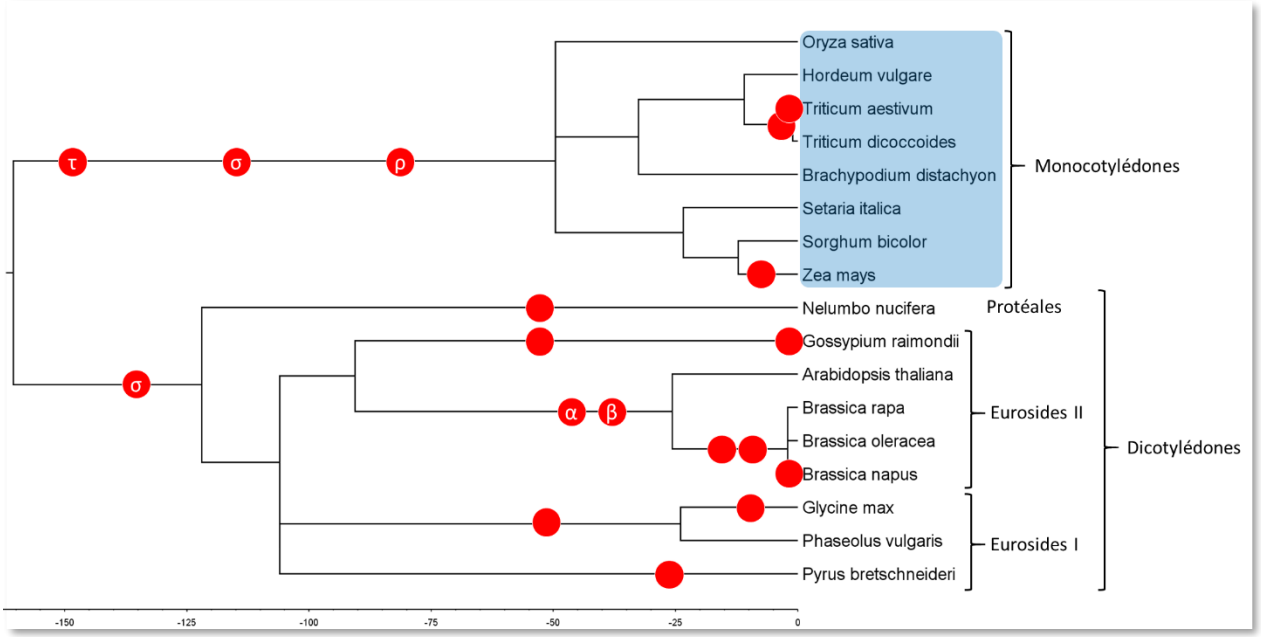


Figure 52 : Phylogénie des espèces incluses dans l'étude bibliographique. Les espèces considérées dans le panel « céréales » sont encadrées en bleu. Les WGD sont figurées par des points rouges.

L'étude de l'évolution d'*Arabidopsis thaliana* a bénéficié de la publication de la séquence du génome au début des années 2000. *A. thaliana* a subi deux évènements de tétraploïdisation α et β sur une période estimée entre 24 et 40 millions d'années (Murat *et al.*, 2015). Le caryotype moderne d'*A. thaliana* compte 5 chromosomes et a été fortement réarrangé post-polyploïdie. *A. Thaliana* a subi un fractionnement biaisé (Thomas *et al.*, 2006) mais, en dépit des pertes de gènes au cours du temps, entre 23% (Blanc *et al.*, 2003) et 29% (Thomas *et al.*, 2006) des gènes demeurent conservées en paires. Une étude récente (Coate *et al.*, 2020) indique que 74% des paires de gènes présentent une expression différentielle dont 3% qui présentent un biais réciproque, c'est-à-dire une seule copie exprimée à la fois, l'une dans un tissu et l'autre dans un deuxième tissu, ce qui constitue une sous-fonctionnalisation effective (Duarte *et al.*, 2006). La comparaison en termes de gènes en paires et singletons indique que les singletons sont plus fortement exprimés que les gènes en paires (Wang *et al.*, 2011). Au sein des paires, la copie dont l'expression est inférieure à celle de son homologue présente en moyenne davantage de substitutions non synonymes, est plus susceptible d'être perdue et a un profil d'expression

différent de celui de ses orthologues dans d'autres espèces (Hoffmann and Palmgren, 2016). Enfin, chez *Arabidopsis thaliana,* les gènes les plus stables, c'est-à-dire ceux qui présentent le moins de mutations, sont les plus méthylés et les plus exprimés (Takuno and Gaut, 2012; Vidalis *et al.*, 2016). Le même lien entre faible taux de mutation et forte méthylation est également retrouvé chez *Brachypodium* et chez le riz (Takuno and Gaut, 2013).

Pour comparer les patrons évolutifs contrastés, le triptyque coton (*Gossypium spp.*), maïs (*Zea mays*) et soja (*Glycine max*) est particulièrement informatif. Ces trois espèces présentent des scénarios évolutifs comparables impliquant une duplication ancestrale et une duplication plus récente, mais des conséquences post-polyploïdie différentes. Le coton a subi une première WGD il y a 60 millions d'années puis une WGD plus récente, il y a 2 millions d'années. Le maïs présente également une paléoduplication il y a 90 millions d'années et une WGD plus récente il y a 5 millions d'années. La situation du soja est relativement similaire : WGD ancestrale il y a 60 millions d'années et seconde WGD il y a entre 5 et 13 millions d'années.

En détail, le coton a subi une paléoduplication il y a environ 60 millions d'années, puis une allopolyploïdisation récente il y a un à deux millions d'années par hybridation de deux génomes (A et D) ayant divergé 5 à 10 millions d'années auparavant. L'étude de l'évènement de paléoduplication met en évidence l'existence de compartiments LF et MF. Au sein de ces deux compartiments, le niveau d'expression moyen des gènes LF est plus élevé que celui des gènes MF. Une hypothèse pour expliquer la sous-expression des gènes en MF est que le compartiment MF présenterait une densité en ET supérieur à celle du compartiment LF, de nature à inhiber l'expression des gènes à proximité. Le compartiment MF présente effectivement une densité d'ET supérieure et davantage de séquences de siRNA à proximité des gènes relativement au compartiment LF, et les gènes MF présentent, en fréquence, plus d'insertions d'ET que les gènes LF. Mais l'étude détaillée du lien entre la présence d'ET à proximité d'un gène et une sous-expression se conclut par une absence de corrélation. Cela semble contredire l'hypothèse indiquant que l'inhibition de l'expression des gènes serait médiée (uniquement) par les ET, et par conséquent qu'ils constitueraient le facteur initiateur de la perte des gènes (Renny-Byfield *et al.*, 2015). Le même auteur complète cette analyse par un second travail portant sur l'expression au sein des paires de gènes issues de la duplication ancestrale. En considérant l'expression dans trois

tissus, les pétales, les feuilles et les graines, de très forts taux d'expression différenciée entre les deux copies sont mis en évidence. Sur 2000 paires de gènes issues de la duplication ancestrale, plus de 99% des paires présentent une différence pour au moins l'un des trois tissus et 93% d'entre elles présentent des différences entre les trois tissus. La différenciation spatiale de l'expression est dans ce cas quasiment généralisée. Pour les graines, la différenciation temporelle a également été analysée à 4 stades de développement, 10, 20, 30 et 40 jours après floraison, *day* post-*anthesis* (*DPA*). Sur 1971 paires de gènes, de façon relativement constante environ 84% (1854 à 1878 paires) des paires présentent des différences pour chacun des quatre stades (Renny-Byfield *et al.*, 2014).

En considérant la même espèce, du point de vue de l'allopolyploïdisation récente ou des polyploïdes synthétiques issus du croisement réalisé entre les sous-génomes parentaux A (*Gossypium arboreum*) et D (*Gossypium herbaceum* et *Gossypium. Raimondii)*, les dynamiques sont différentes dans le cadre des deux croisements (Yoo *et al.*, 2013). Des phénomènes de dominance de niveau d'expression sont détectés dans les deux cas. Chez l'allopolyploïde, le sous-génome A est dominant pour le niveau d'expression à l'échelle du génome, tandis que la tendance est inversée chez certains allopolyploïdes synthétiques. Dans ces cas, il semble que les biais observés reflètent les niveaux relatifs d'expression des gènes des progéniteurs. Sur la base de ces différences initiales, les biais d'expressions semblent ensuite augmenter au cours du temps (au fil des générations) post-polyploïdie. Au niveau de l'expression différentielle des paires, les taux mesurés chez l'allopolyploïde naturel et les allopolyploïdes synthétiques, sont respectivement de 41% et 5%. Ainsi la différenciation de l'expression au sein des paires de gènes apparaîtrait comme un processus progressif au cours du temps. Les hypothèses expliquant ces différences sont, soit la mise en place progressive de différences d'expression, soit des pertes de copies affectant davantage les paires ne présentant pas de différence d'expression que celles présentant des copies différentiellement exprimées, ces dernières devenant par conséquent plus nombreuses en proportion au cours du temps.

Ces tendances observées chez le coton se retrouvent chez le maïs, amplifiées par un intervalle de temps post-polyploïdie plus long. Parallèlement, la structure du caryotype du maïs a été remodelée. Alors que le coton, comme le blé hexaploïde, présente une structure homéologue

(génomes des progéniteurs pas ou peu réarrangés), la tétraploïdisation du maïs remontant à 5 millions d'année a fortement impactée la structure du caryotype (cf. page 80), et des pertes de gènes importantes sont intervenues. Ainsi, Schnable et ses collègues estiment que seules 28 % des paires d'origine ancestrale (5 millions d'années) sont encore des paires aujourd'hui (ce nombre est de 45 % dans le cadre de mon étude). Ces pertes de gènes déséquilibrées entre les segments homéologues définissent un fractionnement biaisé. Les régions dupliquées du maïs ont été inférées par alignement sur le génome du sorgho qui a conservé la structure de l'ancêtre commun aux deux espèces. Pour chaque segment dupliqué chez le maïs, le segment conservant le plus de gènes orthologues au sorgho constitue le sous-génome 1 (LF), l'autre segment constituant alors le sous-génome 2 (MF), (Schnable *et al.*, 2011). Le sous-génome 1 (LF) est plus exprimé et contient, moins d'éléments transposables, et accumule moins de mutations que le sous-génome 2 (MF) (Schnable *et al.*, 2011; Renny-Byfield *et al.*, 2017; Zhao *et al.*, 2017). La définition des sous-génomes relativement aux gènes orthologues est différente de celle des compartiments LF et MF inférés à partir des caryotypes ancestraux reconstruits dans notre étude, bien que les deux méthodes reposent sur le comptage de gènes retenus (ancestraux dans ce travail et orthologues dans le cas de l'étude objet de ce manuscrit). En ce qui concerne les mutations des gènes, le sous-génome 2 (comme le compartiment MF dans mon étude) accumule plus de mutation que le sous-génome 1 (comme le compartiment LF dans mon étude). Ces résultats concordants vont dans le sens de l'hypothèse émise par Freeling *et al.* qui propose que le sous-génome 1 (LF) du maïs subi une sélection purificatrice plus forte que le sous-génome 2 (MF), et que, par conséquent, ce dernier peut donner naissance à des nouvelles variations des caractères phénotypiques (Freeling *et al.*, 2012). La notion de singletons est étudiée mais les comparaisons d'expression sont faites entre singletons du sous-génome 1 et singletons du sous-génome 2 ; dans ce cadre il n'y a pas de différences d'expression détectées entre les singletons (Renny-Byfield *et al.*, 2017). La comparaison entre l'expression des paires et celle des singletons n'est pas étudiée dans les articles auxquels je me suis référé, mais l'expression des paires est analysée. Dans 8 tissus considérés (dont pousse, racine, feuille, tissus vasculaires, épi en développement), il y a entre 55 % et 70 % des paires qui montrent une différence d'expression entre les deux copies (Schnable *et al.*, 2011). Ces chiffres sont légèrement supérieurs mais

demeurent comparables aux 52% de paires différentiellement exprimées chez le maïs que j'ai identifié en considérant les niveaux d'expression au cours du développement du grain. Ces chiffres confirment la persistance au cours du temps des paires de gènes dont les deux copies demeurent co-exprimées chez les céréales. Lorsqu'il y une différence d'expression entre paires, la question de la dominance d'un sous-génome s'exerçant sur le second, reste ouverte. Selon Zhao *et al.* le sous-génome 1 (LF) est le plus exprimé dans 60% des cas et les sous-génome 2 (MF) dans 40% des cas, les proportions variant selon les tissus mais le sous-génome 1 étant toujours le plus exprimé (Zhao *et al.*, 2017), alors que selon Li *et al.* il n'y a pas de dominance d'expression entre les gènes en paires sur les deux sous-génomes (Li *et al.*, 2016). En termes de méthylation des gènes il n'y a pas de différence entre sous-génome (Renny-Byfield *et al.*, 2017). Par ailleurs, l'étude d'une inversion spécifique de certains génotypes issus de populations de maïs adapté aux hauts plateaux mexicains, ne permet pas de détecter d'effets systématique sur l'expression des gènes (Crow *et al.*, 2020).

Le soja a subi un évènement de polyploïdisation il y a 5 à 13 millions d'années. Le caryotype ancestral comprenait 12 protochromosomes pré-polyploïdisation. La WGD a porté ce nombre à 24 protochromosomes qui ont été intensément réarrangés (13 fissions et 14 fusions) pour constituer le caryotype actuel du soja qui compte 20 chromosomes (Murat, Zhang, *et al.*, 2015). En dépit de cette dynamique structurale les pertes de gènes ont été limitées et 80% des paires issues de la dernière WGD demeurent conservées. Ces pertes de gènes ne sont pas biaisées. Il n'y a pas non plus de bais d'expression, ni de contenus en ET, ni de différences de méthylation entre les segments homéologues. Les différences d'expression entre gènes dupliqués sont présentes pour environ 40% des paires sans biais vers l'un ou l'autre segment homéologue quel que soient les tissus analysés (Zhao *et al.*, 2017).

Ainsi le coton, le maïs et le soja présentent 3 cas distincts, respectivement, structure homéologue et dominance des sous-génomes, structure réarrangée et dominance des sous-génomes, structure réarrangée sans dominance des sous génomes. Les hypothèses pour expliquer ces statuts se basent sur les degrés de similarités entre les progéniteurs de ces trois espèces, avec des progéniteurs présentant des différences importantes (niveaux d'expression, méthylation, ET) pour le coton et maïs, et, au contraire, forte similarité (autopolyploïdie) pour le soja. Les résultats

obtenus sur le maïs corroborent ceux présents dans la littérature, à l'exception notable de l'expression dans les compartiments LF et MF. Le comportement du blé se rapproche de celui du coton.

Par ailleurs, une publication (Xu *et al.*, 2018) sur le soja et le haricot traite des différences d'expression et de méthylation au sein des paires de gènes. Cette étude indique que les paires de gènes du soja sont plus exprimées et plus méthylées que les singletons, alors que les tendances sont inversées chez le haricot, les singletons étant plus méthylés et plus exprimés que les paires. Au sein des paires, les niveaux de méthylation des deux copies sont fortement corrélés chez le soja, tandis que chez le haricot ils sont plus divergents. Les résultats du haricot sont identiques à ceux obtenus sur le maïs notamment pour la WGD récente. En termes de fonctions les résultats sont communs aux deux espèces. Les singletons sont enrichis pour des processus liés à la photosynthèse tandis que les paires sont enrichies pour des processus de signalisation.

Les *Brassica* sont également un modèle de choix pour l'étude de la polyploïdie. Les *Brassica* ont divergés à partir d'une triplication ancestrale il y a environ 15 millions d'années, pour aboutir à trois génomes modernes néodiploïdisées : *Brassica rapa* (AA, n=10), *Brassica nigra* (BB, n=8) et *Brassica oleracea* (CC, n=9). Combinés deux à deux, ces trois espèces sont les progéniteurs de trois espèces tétraploïdes : *Brassica juncea* (AABB, n=18), *Brassica carinata* (BBCC, n=17) et *Brassica napus* (AACC, n=19). Ce scénario évolutif est connu sous le nom de triangle de U (Figure 53).

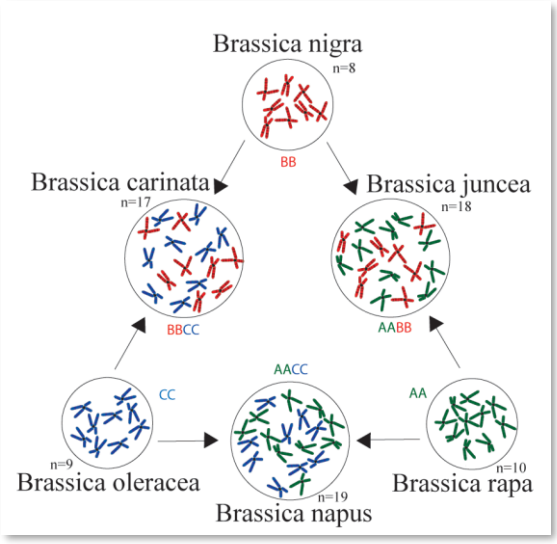


Figure 53 : Triangle de U schématisant les combinaisons entre les trois espèces diploïdes de Brassica à l'origine des trois espèces tétraploïdes.

La disponibilité de l'ensemble des espèces impliquées, la possibilité de générer des néopolyploïdes et la proximité évolutive avec *Arabidopsis thaliana* dont ils ont divergé il y a 27 millions d'années (Murat *et al.*, 2015) font des *Brassica* un modèle particulièrement étudié pour décrypter les conséquences de la polyploïdie.

Le séquençage et l'étude des transcriptomes de progéniteurs diploïdes *Brassica rapa* et *Brassica oleracea* a permis de décrire les effets de la triplication ancestrale (Cheng *et al.*, 2012; Liu *et al.*, 2014; Parkin *et al.*, 2014). La triplication ancestrale est le résultat de deux étapes de polyploïdisation ancestrale impliquant un ancêtre pré-triplication à n=7 chromosomes. Les caryotypes ont subi des réarrangements post-triplication aboutissant aux caryotypes à n=10 et n=9 chromosomes de *Brassica rapa* et *Brassica oleracea*, respectivement. Le dernier génome ancestral hybridé qui a perdu le moins de gènes parmi les 3 sous-génomes agrégés constitue le compartiment LF, les deux autres sous-génomes constituent les compartiments MF1 et MF2. En moyenne 60% des gènes en triplets post-triplication ont été perdus dont 50% des gènes dans le compartiment LF, 65% dans le compartiment MF1 et 70% dans le compartiment MF2. Le réarrangement du caryotype et les pertes de gènes témoignent de la néodiploïdisation intervenue en 15 millions d'années. L'analyse de la régulation de ces compartiments pour les deux espèces aboutit à des résultats convergents. Le compartiment LF est plus exprimé, moins méthylé et accumule moins de SNP que les deux compartiments MF. Le seul élément divergent entre les trois publications est l'analyse du Ks pour lequel deux publications ne mentionnent pas de différences entre compartiments alors que la troisième indique que le compartiment LF accumule davantage de mutations synonymes que les compartiments MF. L'article de Parkin *et al.* examine les triplets post-triplication. Parmi les triplets de gènes conservés, 83% des gènes sont différentiellement exprimés dans la feuille. Au sein des paires, les copies localisées sur le compartiment LF sont plus exprimées que les copies sur MF, mais elles ne montrent pas de différences de méthylation.

L'étude du génome du colza, *Brassica napus*, tétraploïde issu de la fusion de *Brassica rapa* et *Brassica oleracea,* apporte un éclairage supplémentaire sur le comportement des *Brassica* vis-à-vis des évènements de polyploïdisation. La séquence du génome parue en 2014 (Chalhoub *et al.*, 2014) confirme que l'hybridation remonte au maximum à 12 500 ans, c'est donc un évènement

évolutif très récent. Le génome comporte environ 91 000 gènes ce qui correspond globalement à la somme des gènes des progéniteurs. Les sous-génomes An et Cn de *B. napus* sont globalement colinéaires aux génomes de *B. rapa* Ar et *B. oleracea* Co. Ceci comportent 42 320 et 48 847gènes dont 34 255 et 38 661 orthologues des gènes des sous-génomes du colza, respectivement, An et Cn. Le sous génome An correspond donc au sous-génome MF et le sous-génome Cn au sous-génome LF. La même étude indique que 58 % des paires de gènes homéologues ne présentent pas de différences d'expression sur la base des mesures réalisées dans les feuilles et les racines. En observant l'expression dans ces deux tissus, il n'apparait pas de patron évident de dominance en termes d'expression des sous-génomes : les homéologues An ont une expression plus forte que les homéologues Cn dans les racines, le schéma étant inversé dans les feuilles. Une étude parue en 2020, focalisée sur l'étude de l'expression entre paires homéologues dans 4 tissus (tiges, feuilles, fleurs et siliques) permet d'affiner ces résultats (Li *et al.*, 2020). Cette étude indique que pour la majorité des paires (87% en moyenne) les homéologues conservent les profils d'expression de leurs progéniteurs diploïdes. Globalement 78% des paires de gènes présentent un biais d'expression vers le sous-génome A. Les biais varient d'un tissu à l'autre de 70 % et 85 %, les valeurs les plus importantes étant mesurées pour les tiges et les siliques, les plus basses dans les feuilles et les fleurs. Une étude parue en 2021 vient renforcer les conclusions de Li *et al.* et les complète et, en étudiant spécifiquement les paires et les singletons, montre que les singletons sont moins exprimés et plus méthylés que les paires, que les singletons du sous-génomes Cn/LF sont plus méthylés que les singletons du sous-génome An/MF. En ce qui concerne les paires homéologues, cette étude confirme que pour la majorité des paires il n'y a pas de biais d'expression et que les cas d'expressions différentielles se partagent equitablement entre les deux sous-génomes, et dans tous les cas de figure, les copies du sous-génome Cn/LF sont les plus méthylées (Zhang *et al.*, 2021). Ces résultats montrent l'impact de l'état pré-polyploïdie des progéniteurs. Ils sont globalement cohérents avec ceux résutats obtenus dans le cadre du travail présenté ici, hormis en ce qui concernent les différences d'expression entre gène homéologues. Sur ce point il est possible que le temps écoulé post-polyploïdie soit trop faible pour que des différences se mettent en place dans le cas de l'allopolyploidie du colza datant de 12 500 ans.

A l'instar du colza, le génome du blé hexaploïde est également un modèle très étudié de polyploïdisation récente. Son génome est organisé en 3 sous-génomes homéologues ayant conservés la structure commune à leurs ancêtres diploïdes comportant de 7 chromosomes. Il n'a pas subi de fusion ni de fission suite aux deux évènements de polyploïdisation récents (cf. page 80). Bien que les caryotypes ancestraux soient maintenus post-polyploïdie, la méiose est fonctionnelle grâce au contrôle par le locus *Ph1* de l'appariement entre chromosomes homéologues (Riley and Chapman, 1958). Alors que les structures de chromosomes sont conservées, des changements entre sous-génomes ont opéré aux niveaux des gènes en termes structural (PAV), contrairement au génome du colza, et au niveau des régulations. Ainsi seulement 47% des groupes d'homéologues correspondent à des triplets avec une seule copie de gène par sous-génome (Appels *et al.*, 2018). Plusieurs publications traitent de l'étude des niveaux d'expression en fonction du statut des gènes homéologues (en triplets, en paires ou en singletons). Elles indiquent que 45% (Mutti *et al.*, 2017) à 80% (Juery *et al.*, 2020) des gènes homéologues conservés en triplets ne présentent pas de différence d'expression. De plus, selon ce dernier travail, si les homéologues sont conservés en paires et non en triplets, ce qui implique la perte d'une copie, le taux de gènes ne présentant pas de différence d'expression baisse à 64%. Par ailleurs, les paires et les singletons seraient principalement situés dans les régions télomèriques et présenteraient des niveaux d'expression plus faibles que les triplets. Les paires, relativement aux triplets, seraient enrichies en fonctions liées à l'adaptation. Une troisième étude (N., Zhao *et al.*, 2020) qui compare les niveaux d'expression entre les paires de gènes homéologues appartenant aux génomes A et B montre que 70% des paires ne présentent pas de différence d'expression, et que les paires différentiellement exprimées cumulent davantage de mutations, Ka et Ks, que celles qui ne présentent pas de changements d'expression. L'ensemble de ces données tendent à confirmer mes résultats confirmant qu'environ la moitié des paires, ou triplets, présentent le même niveaux d'expression entre copies. Il est à noter que les questions des différences entre les compartiments LF et MF, entre les paires (ou triplets) et les singletons ne sont pas traitées dans les publications considérées.

Globalement la conservation des caryotypes homéologues du blé et du colza suite à des évènements de polyploïdisation récents (≥ 1 million d'années) contraste avec les caryotypes

réarrangé du maïs (ou encore du soja) et soulève la question du chemin parcouru par le maïs tétraploïde ancestral (après la WGD datant de 5 millions d'années) : est-t-il passé par un caryotype non-réarrangé supposant l'existence d'un système de contrôle des appariements chromosomiques, ou a-t-il subi de rapides réarrangements post-polyploïdisation ? Et, en corollaire, dans quel mesure l'évolution future du blé va-t-elle se faire par des réarrangements chromosomiques ? Une partie des réponses à ces questions complexes est apportée en étudiant les perturbations de la méiose par la colchicine chez différentes espèces de maïs (Poggio and González, 2018). Chez les plantes, la colchicine appliquée au début de la méiose supprime l'effet des mécanismes régulant les appariements homéologues, ce qui entraine la formation de multivalents. Appliqué sur le maïs, le traitement à la colchicine entraine la formation de quadrivalents, fruits de l'appariement entre chromosomes homéologues. Ces résultats suggèrent la présence d'un locus régulant ces appariements chez le maïs (*PrZ*), homologue en termes de fonction au locus *Ph1* du blé. La divergence en termes de structure des sous-génomes (réarrangements chromosomiques et mutations) pourrait ne pas être suffisante pour assurer une méiose régulière. La divergence et le contrôle des appariements seraient des systèmes indépendants mais complémentaires pouvant agir conjointement dans le même noyau.

Outre l'observation des variations en termes d'évolution des caryotypes et les liens entre les compartiments LF et MF et les caractéristiques (mutations/méthylation/expression), mon étude met en évidence, d'une part, la différence de comportements des gènes qu'ils soient conservés en paires ou qu'ils le soient en singletons, et, d'autre part, des patrons de liens entre méthylation et expression des gènes. La bibliographie est limitée sur le premier point et plus vaste sur le second mais le lien entre méthylation et expression demeure non élucidé.

En ce qui concerne les différences entre le comportement des paires et de singletons, l'étude du génome du poirier, *Pyrus bretschneideri*, apporte une vision complète et détaillée (Q., Li *et al.*, 2019). Le poirier a subi une duplication il y a 30 millions d'années qui n'a pas donné lieu à une dominance d'un sous-génome. L'étude des blocs homologues ne montre pas de différence en termes de taux de pertes de gènes, de mutations, d'expression ou de méthylation. 54% des paires de gènes issues de la WGD sont différentiellement exprimées. La comparaison des singletons et des paires indique que les singletons sont plus fortement exprimés et dans davantage de tissus

différents que les paires. Les singletons sont plus méthylées en CG au niveau des promoteurs que les paires, alors qu'il n'y a pas de différence révélée au niveau de la méthylation du gène (*gene body*). En CHG la tendance est la même au niveau des promoteurs et des gènes eux-mêmes, les singletons apparaissant plus méthylés que les paires. En termes de mutations, il n'y a pas de différence en termes de Ks mais en termes de Ka et de Ka/Ks les singletons présentent de plus fortes valeurs que les paires.

Une deuxième étude portant sur le lotus d'orient, *Nelumbo nucifera*, appartenant à l'ordre des *Nymphaeales*, apporte également des réponses sur la question des singletons. *N. nucifera* a subi une seule paléoduplication datée de 60 millions d'années. Dans un travail qui traite spécifiquement de l'évolution des gènes dupliqués post-polyploïdie les auteurs montrent que, d'une part, au sein des paires, les copies du compartiment MF sont plus méthylées et plus exprimées que les copies du compartiment LF, et que, d'autre part, les singletons sont plus fortement exprimés et dans davantage de tissus différents que les paires. De plus, ils constatent également que, relativement aux gènes conservés en paires, les singletons sont plus méthylés (*gene body*) et accumulent moins de mutations (Shi *et al.*, 2020). Ces conclusions sont en accord en ce qui concerne les singletons par rapport aux paires, avec une revue de De Smet *et al.* qui, en compilant les résultats sur 20 espèces d'angiospermes, indique que les singletons sont effectivement plus conservés (présentant moins de mutations), plus méthylés et plus exprimés que les paires. Par ailleurs cette revue montre que des singletons sont maintenus en copies uniques au fil de plusieurs WGD successives et qu'il y aurait donc une sélection active contre la redondance de certains gènes (DeSmet *et al.*, 2013). Le patron retrouvé systématiquement dans ces travaux à savoir hyper-méthylation et surexpression des singletons interroge évidemment le lien entre les deux niveaux de régulation. Wang *et al.* étudient spécifiquement la méthylation en CG des paires de gènes chez le riz et montrent que les gènes présentant des différences de niveaux de méthylation sont ceux qui présentent des différences d'expression (Wang *et al.*, 2017).

L'ensemble des données issues de cette revue bibliographique permet, sous forme d'une matrice d'informations, d'observer les convergences et les divergences entre les résultats publiés et les conclusions de cette étude (Figure 54).

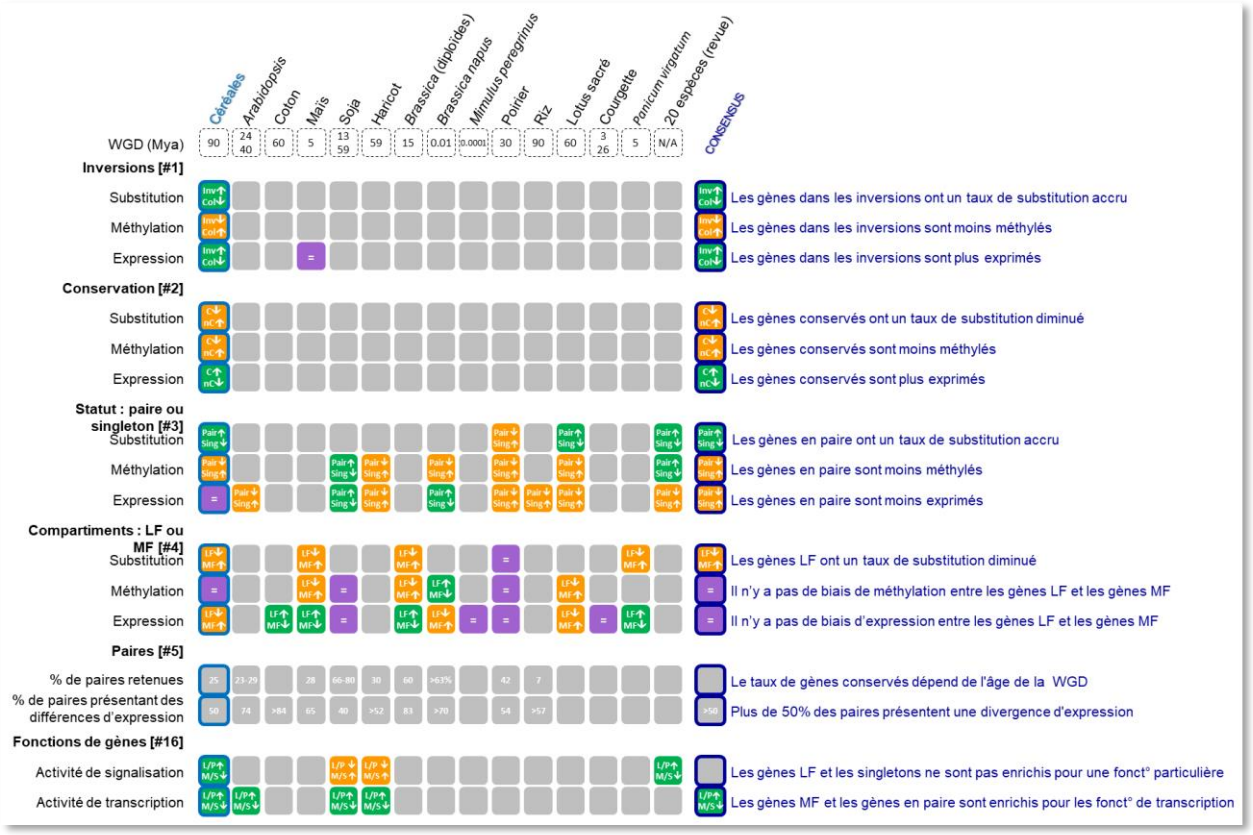


Figure 54 : Comparaison entre les tendances génériques identifiées sur les céréales et les conclusions issues de la bibliographie. Pour chaque espèce (et une revue) en colonne, les tendances sont indiquées par un code couleur. En vert quand le premier terme présente des valeurs significativement supérieures au second. En orange, dans le cas contraire. En violet, en l'absence de différence. En gris, lorsqu'il n'y a pas d'informations sur un cas de figure dans la bibliographie. Dans le volet « paires » sont indiqués les pourcentages de paires ancestrales retenues et de paires dont les deux gènes présentent des différences d'expression. La colonne 'consensus' résume les conclusions issues de la présente analyse et des informations de la bibliographie pour chacun des phénomènes considérés.

Globalement les résultats obtenus trouvent des confirmations dans la bibliographie. Cependant, des nuances existent, elles sont abordées ici, point par point.

La comparaison entre singletons et paires est utilisée relativement rarement pour caractériser l'évolution post-polyploïdie mais globalement dans les exemples recensés, les conclusions issues de mes observations, à savoir les singletons présentant moins de mutations et étant plus méthylés que les paires, sont confirmées. Alors que je ne trouve pas globalement de différences d'expression entre singletons et paires (hormis chez le maïs, avec les singletons plus exprimés que

les paires), la tendance majoritaire dans la bibliographie est que les singletons sont plus exprimés que les paires. En ce qui concerne les compartiments LF et MF, les analyses des mutations et de la méthylation vont dans le sens de mes résultats. Par contre, au sein des compartiments LF et MF, les données issues de la revue bibliographique sont globalement opposées à mes conclusions en termes d'expression. Cependant, les résultats chez le colza et chez le lotus sont concordants avec l'étude présentée ici. De plus, il faut garder à l'esprit que ces résultats reposent sur l'analyse des compartiments LF et MF issus de la duplication ancestrale ρ pour le blé et *Brachypodium* et que cette paléoduplication et la duplication récente pour le maïs sont dissociées, tandis que la majorité des études considèrent les effets cumulés de plusieurs WGD.

Concernant l'expression des gènes en paires, dans la plupart des études les pourcentages de paires de gènes différentiellement exprimés, pour la majorité des espèces plus de 65% des paires, sont supérieurs à ceux que j'ai constatés sur le maïs, le blé et *Brachypodium* où les DEG entre les trois stades de développement de la graine représentent 50% des paires. Les différences sont à pondérer par le fait que les seuils considérés peuvent être différents. Par exemple chez le coton 84 à 99% des paires sont considérées comme différentiellement exprimés, DEG *Differentially Expressed Genes*, sur la base d'un facteur de 1,5 de différence d'expression entre les deux copies d'une paire. Dans mon travail, le seuil était plus strict, fixé à 2. Cependant, même en considérant un facteur 2 entre les DEG, plus de 80% des paires de gènes issues de la duplication ancestrale du coton présentent des différences d'expression entre tissus. Mais au-delà des chiffres, il semble que la principale conclusion est la même dans le cas des polyploïdisations anciennes, à savoir que des gènes dupliqués demeurent coexprimés, autrement dit, qu'il y a toujours une proportion de paires que ne sont ni néo- ni sous-fonctionnalisées.

Les résultats sur les fonctions des gènes enrichies en fonction de leur statut ou du compartiment auquel ils appartiennent sont généralement conformes à mes conclusions, en particulier pour les paires et les gènes du compartiment MF. Pour les singletons et les gènes du compartiment LF, les résultats de De Smet *et al.* (2013) portant 20 espèces d'angiospermes sont concordants avec mes conclusions, à savoir un enrichissement pour les processus de signalisation, tandis que le soja et le haricot montrent un enrichissement pour les processus liés à la photosynthèse.

La majorité des études établit qu'il y a un lien entre méthylation et expression des plantes sans que ce lien soit quantitativement exploré. Je présente une vision basée sur des seuils de méthylation à la fois pour la méthylation des promoteurs et pour la méthylation *gene body*, dans les deux cas les valeurs de méthylation les plus élevées affectant négativement l'expression.

4.5 Proposition d'un modèle évolutif

Sur la base des résultats du travail présenté dans ce manuscrit, confrontés aux informations disponibles dans la bibliographie, je présente les 10 conclusions majeures sur la structure et la régulation des génomes des angiospermes post-polyploïdie. Dans la majorité des cas, les conclusions issues de mes résultats ont été confirmées dans la bibliographie. Elles sont alors précédées du symbole '➔' et ceux sont ces conclusions qui sont retenues pour proposer un modèle général. Dans une minorité des cas, elles sont différentes des conclusions globalement présentées dans les publications analysées. Ces conclusions sont précédées du symbole '🖹'. Enfin, certains aspects ne sont pas traités dans les publications considérées. Les conclusions se basent exclusivement alors sur les résultats présentés ici et sont précédées du symbole '➔'.

Les conclusions majoritairement observées dans nos travaux et les données bibliographiques sont :

① ➔ Les génomes des angiospermes ont évolué par le biais de fusions et de fissions chromosomiques, sans réutilisation des points de fusions. *Chez les graminées, en particulier, les génomes ont évolué à partir d'un ancêtre de 7 chromosomes avec 16K gènes conservés ancestralement. Ces gènes correspondent à des fonctions cellulaires de base. Les événements d'inversion ont tendance à se produire plus fréquemment dans un contexte de polyploïdie que chez les diploïdes.*

② ➔ Les gènes conservés sont moins méthylés (promoteurs et *gene body*) et plus exprimés que les gènes spécifiques.

③ ➔ Les gènes des régions inversées sont moins méthylés (promoteurs) et plus exprimés que les gènes des régions conservant l'orientation ancestrale.

④ 🗎 Les singletons présentent moins de mutations (plus stables) et sont plus méthylés. Dans la bibliographie et dans des observations sur le maïs (pas de différence chez *Brachypodium* et le blé), les singletons sont plus exprimés que les gènes en paires.

⑤ ➔ Les gènes LF présentent moins de mutations (plus stables) que les gènes MF.

⑥ 🗎 Il n'y a pas de biais de méthylation, ni d'expression entre les gènes LF et MF

⑦ ➔ La co-expression des deux copies issues d'une paire de gènes ancestrale peut perdurer.

⑧ ➔ Il y a une corrélation négative entre la méthylation de l'ADN et l'expression des gènes pour un sous-ensemble de gènes présentant des niveaux de méthylation extrêmes. Cette conclusion soutient l'hypothèse que la méthylation de l'ADN joue un rôle dans le contrôle de l'expression des gènes. Je montre que ce contrôle n'est pas linéairement corrélé aux valeurs de méthylation, mais s'exerce en fonction de seuils de densité de méthylation en CG.

⑨ 🗎 Les singletons et les gènes LF ne sont pas enrichis pour une fonction particulière

⑩ ➔ Les paires et les gènes MF sont enrichis pour les mêmes fonctions, notamment les activités de transcription.

Je propose à partir de ces conclusions, un modèle générique d'évolution des génomes post-polyploïdie qui illustre comment les conséquences de la polyploïdisation (perte de gènes, mutation, changements de l'expression et de la méthylation) ont façonné et permis le succès évolutif de multiples espèces de graminées depuis 100 millions d'années. (Figure 55).

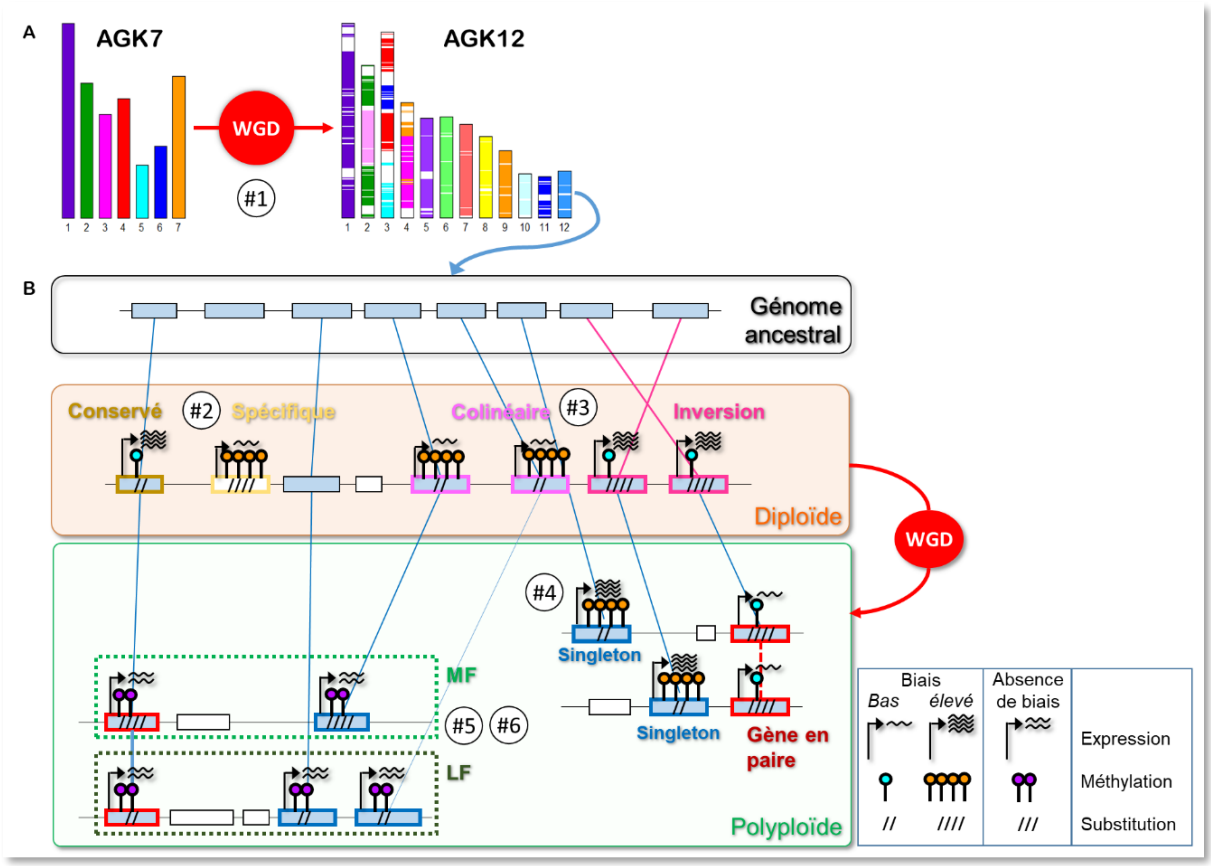


Figure 55 : Modèle générique d'évolution des génomes d'angiospermes post-polyploïdie. A : WGD ancestrale aboutissant à un caryotype réarrangé (illustrée ici par la transition d'AGK7 vers AGK12 des Triticeae). B : néodiploïdisation d'AGK12 donnant l'espèce 1, caractérisée par la perte de gènes ancestraux (en bleu) et le gain de gènes spécifiques en blanc. C. Espèce 2, polyploïde résultant de la WGD de l'espèce 1. Les paires de gènes sont encadrées en rouge et les singletons en bleu. Les compartiments LF et MF sont encadrés en vert foncé et vert clair. Les numéros renvoient aux conclusions développées ci-dessus

Notre étude actuelle démontre que les espèces polyploïdes partageant un événement WGD commun présentent souvent les mêmes profils de reprogrammation de la structure et de la régulation des génomes et de dynamique évolutive. Ces modèles sont cependant difficiles à généraliser à travers des événements WGD indépendants en raison de facteurs qui ne sont pas directement reliés à la WGD, tels que la sélection et la domestication des cultures. La polyploïdie est liée sans équivoque au succès évolutif des graminées durant les 100 derniers millions d'années, bien qu'il reste délicat d'attribuer ce succès à des conséquences génomiques particulières de la polyploïdisation, suggérant plutôt que les polyploïdes exploitent le potentiel de duplication du génome, au moins partiellement, de manière spécifique pour chaque espèce.

Dans l'ensemble, l'étude actuelle démontre explicitement que la reprogrammation post-polyploïdie est plus complexe que ce qui est traditionnellement rapporté dans la littérature dans le cadre de l'étude d'une espèce et appelle à une comparaison critique et complète entre espèces issues d'évènements de polyploïdisation indépendants, même au-delà des graminées au centre de cette étude. Les progrès des technologies de séquençage permettant l'accès à près de 100 génomes de plantes, associés à des données omiques massivement disponibles dans le domaine public avec de nombreuses conditions expérimentales (différents tissus, différentes contraintes biotiques et abiotiques…), ouvrent désormais la porte à une analyse omique comparative à large échelle pour décoder les mécanismes de la reprogrammation génomique post-polyploïdie à l'échelle des angiospermes.

III. CONCLUSIONS ET PERSPECTIVES

1 Conclusions

Ce travail avait pour but de caractériser les conséquences de la polyploïdisation sur la structure et la régulation des génomes des graminées. Afin de disposer de données comparables entre elles pour identifier les tendances génériques, un panel de 8 espèces a été étudié en mettant en œuvre les mêmes méthodes pour définir les compartiments géniques et les statuts des gènes, et pour comparer en fonction de ces caractéristiques, les données traduisant les dynamiques structurales (substitution, Ka, Ks) et les régulations de l'expression et de la méthylation. L'analyse des pertes de gènes par rapport aux caryotypes ancestraux AGK7 et AGK12, qui définissent les compartiments LF et MF, a mis en évidence le fractionnement biaisé qui affecte l'ensemble des génomes considérés suite à la paléoduplication, et le maïs après la WDG spécifique datée de 5 millions d'années, mais qui est absent entre les sous-génomes du blé hexaploïde assemblés lors de deux évènements de polyploïdisation récents. L'analyse des GO des compartiments LF et MF et des gènes en singletons ou en paires des gènes a montré que les gènes sont retenus de façon déterminée par leurs fonctions et non pas aléatoirement.

Ces résultats confirment, et complètent en prenant en compte le compartiment LF/MF, des analyses précédemment menées sur plusieurs espèces de *Brassicaceae* polyploïdes présentant des WGD indépendantes, qui montrent que les gènes retenus en paires quelles que soient les espèces considérées ont des enrichissements similaires en termes de GO (Mandáková *et al.*, 2017). Ainsi certaines fonctions détermineraient la rétention des gènes en paires. L'hypothèse expliquant ce patron de rétention des paires est l'effet dose. L'effet dose correspond au maintien du niveau d'expression des deux copies dupliquées chez le polyploïde au même niveau que celui du gène en copie unique chez l'ancêtre diploïde, conduisant à un doublement de l'expression au niveau de la cellule entière. Cette « surdose » ayant pour conséquence d'augmenter l'efficacité de certaines voies métaboliques (Guo *et al.*, 1996; Birchler and Veitia, 2019). L'effet dose concernerait les gènes au cœur de voies métaboliques qui présentent de nombreuses connexions, les gènes appartenant à des voies sensibles à la stœchiométrie. Ces gènes, à forte connexion au sein des réseaux métaboliques, seraient plus fortement conservés après la

polyploïdie (Birchler and Veitia, 2012). En outre, il apparait dans la bibliographie que les paires dont les deux copies sont exprimées au même niveau présentent moins de mutations que les paires dont les copies sont différentiellement exprimées. Cette tendance va dans le sens d'une conservation des gènes dans leur état initial post-polyploïdie dès lors que leur expression conjointe se maintient après l'événement de duplication du génome. Ce constat va dans le sens de l'hypothèse de l'effet dose. En effet, considérant que le maintien de l'expression des deux gènes procure un avantage au polyploïde, une pression de sélection s'applique sur la séquence des gènes et sur leur régulation, pour conserver une production de protéines adéquate en termes de conformité des peptides codés et de quantité. Cependant, et c'est sans doute les principaux enseignements de ce travail, les mécanismes à l'œuvre post-polyploïdie sont toujours multiples et interconnectés, et l'effet dose ne peut être le seul phénomène à prendre en compte pour expliquer la rétention des gènes post-polyploïdie (Conant *et al.*, 2014). Il convient de considérer également les effets de la hausse de diversité génétique conférée par la présence des copies homéologues. Deux génotypes qui se combinent au sein d'un néopolyploïde peuvent avoir été soumis au cours de leurs évolutions à des contraintes populationnelles, biotiques ou abiotiques, divergentes et, par conséquent, avoir développé des versions alléliques différentes à partir de gènes ancestraux communs. Lorsque ces génotypes se combinent au sein d'une même cellule après la polyploïdisation, ces différences entre les copies homéologues en présence constituent une source potentielle d'interactions nouvelles entre allèles à l'origine d'effets transgressifs positifs. Ces effets s'expliquent par l'effet soit de la compensation de mutations délétères, soit de la complémentarité des allèles. Dans les deux cas, ils peuvent permettre à la plante de faire face à un panel élargi de pressions de sélections (Washburn and Birchler, 2014). Ces effets s'apparentent à la notion d'hétérosis dans le champ de la sélection variétale. Au hasard des apparitions d'individus polyploïdes (cf. page 52), certaines combinaisons pourront dans une situation donnée apporter une réponse à une pression de sélection particulière ou conférer un avantage compétitif par rapport au reste de la population. Contrairement à l'effet dose, la combinaison d'allèles induirait la conservation de paires de gènes en réponse à des conditions particulières de l'environnement, plutôt qu'à des nécessités endogènes de la plante, générant ainsi des patrons de rétention de gènes différents entre espèces, et au fil des WGD, en

comparaison de ceux dus à l'effet dose. Ainsi sur la base de leurs fonctions certaines paires de gènes seraient en quelque sorte à l'abri des pertes de copies post-polyploïdie tandis que d'autres paires de gènes seraient davantage susceptibles de perdre une copie. Une étude sur *Tragopogon miscellus*, une plante de la famille des astéracées, montre que les patrons de rétention et pertes de copies se mettent rapidement en place, dans ce cas en l'espace de 40 générations post-polyploïdie, montrant, par ailleurs, des similitudes avec le patron de conservations des gènes identifié suite la paléoduplication spécifique des astéracées (Buggs *et al.*, 2012).

Considérant que des paires sont conservées, il convient de s'intéresser aux paires dont une copie n'est pas, ou faiblement, exprimée, sujette aux mutations ou tout simplement éliminée. Pourquoi ces pertes d'expression ou pertes de gènes affectent-elles préférentiellement un sous-génome dans le cas de fractionnement biaisé, ou, au contraire se font de façon équilibrée entre sous-génomes ? L'hypothèse admise est que le fractionnement biaisé est le résultat des différences préétablies entre les génomes des progéniteurs dans le cas de l'allopolyploïdie, les autopolyploïdes n'ayant par nature pas de différences, présentant un fractionnement neutre (Garsmeur *et al.*, 2014). Mes résultats montrent un fractionnement biaisé définissant des compartiments LF et MF post-polyploïdisation ancestrale, mais ils montrent qu'il n'y a pas de différence de méthylation entre les gènes des deux compartiments dans les espèces modernes. Ils confortent le postulat que l'impact de la méthylation ne s'exerce pas sur la base de différences locales de méthylation entre copies de gènes homéologues mais, est plutôt le résultat de différences de méthylation entre chromosomes homéologues à l'échelle des grandes régions du génome, voire de chromosomes entiers. Ceci implique non seulement la méthylation des gènes, mais également et peut-être essentiellement celle des régions répétées, plus méthylées et représentant une plus forte proportion du génome que les gènes (Chen *et al.*, 2015). Cette hypothèse va dans le sens de l'absence de lien direct la présence d'ET méthylés à proximité des gènes et leur expression émise par Renny-Byfield, précédemment décrit (cf. page 128). L'exploration d'autres types de modifications épigénétiques, chromatine et histones notamment, permettra de faire avancer la connaissance sur l'importance des différences entre les sous-génomes des progéniteurs des polyploïdes.

Quelle que soit l'origine des différences entre les compartiments, il apparait clairement que les gènes des compartiments LF sont globalement plus stable, présentant moins de mutations, tant en considérant l'ensemble les gènes qu'en se focalisant sur les gènes en paires. Le fait de présenter davantage de mutations confère au compartiment MF, et particulièrement aux gènes conservés en paires, un rôle de réservoir et même de création de diversité les gènes pouvant plus librement évoluer par sous- ou néo-fonctionnalisation. Mais d'après mes résultats, il apparait clairement que le compartiment MF est bien moteur dans l'expression globale du génome, ses gènes étant en moyenne plus exprimés que ceux, par ailleurs, plus nombreux du compartiment LF. Il serait pertinent de vérifier si cette tendance se retrouve en multipliant les tissus considérés pour mesurer l'expression. Si cette tendance se trouvait confirmée, il serait informatif d'identifier si parmi les gènes les plus exprimés se trouvent des copies de gènes néo- ou sous-fonctionnalisés, ce qui constituerait un indice du succès évolutif de fonctions acquises par le polyploïde.

Parallèlement à l'analyse de la structure et de la régulation entre compartiments, je me suis intéressé aux impacts du statut des gènes qu'ils soient conservés en paires ou présents en singletons. Les singletons, accumulant moins de mutations, et étant davantage méthylés et exprimés que les paires, présentent des caractéristiques propres et constituent un statut à considérer pour étudier les impacts de la polyploïdie. Pour aller plus loin dans l'étude du comportement des singletons, il serait pertinent d'adopter une définition commune pour les singletons. Dans mon étude, ils sont définis relativement aux gènes ancestraux pré-WGD dont on sait qu'ils ont été dupliqués. En appliquant cette définition, il deviendrait possible de comparer les trajectoires évolutives de diverses espèces pour identifier des similarités et des spécificités à l'image du travail réalisé sur la rétention des gènes en paires.

Ainsi le travail qui vient d'être présenté apporte une vision globale et intégrée des phénomènes activés post-polyploïdie tout en soulevant de nouvelles hypothèses qui pourront être testées grâce aux avancées des technologies de séquençage et d'analyses -omiques.

2 Perspectives

La question de la compréhension des mécanismes activés post-polyploïdie trouve de plus en plus de réponses, mais de nombreuses hypothèses restent à vérifier et le puzzle demeure incomplet. Les connaissances actuelles s'appuient majoritairement sur la caractérisation d'évènements passés par un nombre limité de descripteurs de la structure et de la régulation. Pour ajouter les pièces manquantes deux voies semblent particulièrement prometteuses : d'une part, augmenter le nombre de polymorphismes structuraux caractérisés chez les polyploïdes et, d'autre part, tirer parti de la possibilité de produire des polyploïdes synthétiques.

En s'appuyant sur des assemblages de génomes de haute qualité, il est désormais possible de prendre en compte de façon exhaustive tous les types de mutations structurales (SNP, CNV, insertion de ET, insertion/délétion, translocations, inversions, fusion/fission et modifications des centromères) et épigénétiques (modifications de la méthylation des nucléotides, mais aussi modifications des histones et de l'état chromatinien) susceptibles d'influer sur la régulation des gènes et leur évolution à court, moyen et long terme (Berdan *et al.*, 2021). L'analyse des inversions abordées dans ce travail, à l'image de l'ensemble des variations structurales, apparait comme un nouveau champ de recherche, car les grandes variations structurales sont susceptibles d'avoir un impact sur l'expression des gènes et la recombinaison. Leur nombre et leur ampleur au niveau interspécifique, mais aussi intraspécifique demande à être évaluée. En effet, notre analyse comparative des 8 espèces repose sur la comparaison des 8 génomes de référence à disposition à ce jour. Passer à l'analyse de 8 pangénomes, intégrant l'ensemble des variations structurales à l'échelle populationnelle, permettra une analyse plus exhaustive de la reprogrammation génomique post-polyploïdie. Toutefois, cela implique de produire des pangénomes pour un ensemble d'espèces et de développer les méthodes et outils nécessaires à la comparaison de pangénomes entre espèces. La réalisation de ce travail constitue sans nul doute l'avenir des travaux de génomique comparative à court terme.

Les polyploïdes synthétiques constituent le modèle de choix pour étudier les phénomènes de régulation post-polyploïdie en recréant les évènements du passé. Des polyploïdes synthétiques de blé ont permis d'étudier le maintien des structures homéologues et de montrer que les pertes

de gènes ou la modification de l'expression de ceux-ci, ne se font pas de façons aléatoires (A., Li *et al.*, 2015). De nouvelles études sur les polyploïdes synthétiques, sur plusieurs générations post-polyploïdie, permettront de tester les hypothèses sur l'impact des éléments transposables, de la méthylation, de l'accessibilité de la chromatine sur la reprogrammation génomique post-polyploïdie.

Notre travail montre que la reprogrammation génomique post-polyploïdie est source d'une plasticité structurale et fonctionnelle à l'origine, potentiellement, de nouveaux phénotypes. Une parfaite illustration se trouve dans l'adaptation d'une variété polyploïde de la brassicacée *Pugionium* aux contraintes du milieu désertique. Le génotype polyploïde présente une expansion post-polyploïdie des familles de gènes liées aux réponses au stress abiotique et à la biosynthèse de la lignine qui lui confère un potentiel accru d'adaptation aux conditions environnementales extrêmes du milieu aride où elle s'est implantée (Hu *et al.*, 2021). Cet exemple, parmi d'autres, pousse à s'interroger sur la possibilité d'exploiter la plasticité génomique post-polyploïdie en sélection variétale. Ainsi quelques espèces polyploïdes synthétiques se sont imposées comme espèces cultivées telles que la pastèque triploïde (*Citrullus vulgaris* Schard.), la betterave triploïde (*Beta vulgaris* L.), le kiwi tétraploïde (*Actinidia chinensis*), le pommier tétraploïde (*Malus* ssp. Mill) ou le bananier banana (*Musa* ssp.) (Ruiz *et al.*, 2020). L'augmentation de la synthèse de molécules d'intérêt chez les versions polyploïdes, naturelles ou synthétiques, de plantes aromatiques et médicinales constitue une application potentielle de la polyploïdie pour la sélection (Iannicelli *et al.*, 2020). De même, la production de polyploïdes synthétiques apparaît comme une solution efficace pour optimiser la production de latex (Luo *et al.*, 2018). Outre la valeur intrinsèque des polyploïdes, le passage à l'état polyploïde permet de restaurer la fertilité de plante hybride. Cette possibilité de générer des allopolyploïdes synthétiques à partir d'un hybride diploïde est particulièrement utile et prometteuse en sélection également (Meeus *et al.*, 2020; Koide *et al.*, 2020).

La polyploïdie est incontestablement une force majeure pour l'évolution des plantes à fleur. L'accumulation des connaissances et les avancées des méthodes d'analyse des génomes permettent d'accumuler des données pour comprendre et, de plus en plus, modéliser les effets de la programmation génomique post-polyploïdie qui serait à l'origine de cet avantage

adaptatif des espèces polyploïdes. Même si les preuves apportées durant ce travail sur les modifications de la structure et de la régulation des génomes après un événement de polyploïdie sont évidentes, l'absence d'une complète généricité des réponses génomiques, pouvant varier selon l'événement de polyploïde étudié ou de l'espèce considérée, compliquent l'établissement d'un modèle unique et la proposition d'un seul mécanisme qui pourrait être à l'origine de cette reprogrammation génomique post-polyploïdie. Toutefois, grâce à son potentiel unique pour produire de la diversité génétique et générer de la nouveauté en termes de régulation, la polyploïdie apparaît comme une solution à fort potentiel pour sélectionner de nouvelles variétés de plantes contribuant à développer une agriculture répondant aux défis des changements globaux, climatiques et sociétaux.

IV. BIBLIOGRAPHIE

Acharya, D. and Ghosh, T.C. (2016) Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics*, **17**, 1–14.

Adams, C.P. and Kron, S.J. (1997) Method for performing amplification of nucleic acid with two primers bound to a single solid support. Patent US5641658.

Adams, M.D., Kelley, J.M., Gocayne, J.D., et al. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Alekseyev, M.A. and Pevzner, P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, **19**, 943–957.

Alger, E.I. and Edger, P.P. (2020) One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.*, **54**, 108–113.

Aliscioni, S., Bell, H.L., Besnard, G., et al. (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C 4 origins. *New Phytol.*, **193**, 304–312.

Alonge, M., Wang, X., Benoit, M., et al. (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145-161.e23.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.

Anssour, S. and Baldwin, I.T. (2010) Variation in Antiherbivore Defense Responses in Synthetic Nicotiana Allopolyploids Correlates with Changes in Uniparental Patterns of. , **153**, 1907–1918.

Appels, R., Eversole, K., Feuillet, C., et al. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.

Arkhipova, I.R. (2018) Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol. Biol. Evol.*, **35**, 1332–1337.

Arrigo, N. and Barker, M.S. (2012) Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.*, **15**, 140–146.

Aury, J., Jaillon, O., Duret, L. and Al., E. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**, 171–178.

Aury, J.M., Engelen, S., Istace, B., et al. (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *Gigascience*, **11**, 1–18.

Avery, O.T., Macleod, C.M. and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, **79**, 137–158.

Avni, R., Nave, M., Barad, O., et al. (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, **97**, 93–97.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K. and Colot, V. (2019) Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nat. Commun.*, **10**.

Baerdemaeker, N.J.F. De, Hias, N., Bulcke, J. Van den, Keulemans, W. and Steppe, K. (2018) The effect of polyploidization on tree hydraulic functioning. *Am. J. Bot.*, **105**, 161–171.

Baidouri, M. El, Murat, F., Veyssiere, M., et al. (2017) Reconciling the evolutionary origin of bread wheat

(Triticum aestivum). *New Phytol.*, **213**, 1477–1486.

Bao, Y., Hu, G., Grover, C.E., Conover, J., Yuan, D. and Wendel, J.F. (2019) Unraveling cis and trans regulatory evolution during cotton domestication. *Nat. Commun.*, **10**, 1–12.

Barwell, J.G., O'Sullivan, R.B., Mansbridge, L.K., Lowry, J.M. and Dorkins, H.R. (2018) Challenges in implementing genomic medicine: the 100,000 Genomes Project. *J. Transl. Genet. Genomics*, **2**, 1–10.

Bashir, A., Volik, S., Collins, C., Bafna, V. and Raphael, B.J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.*, **4**, 1–14.

Beest, M. Te, Roux, J.J. Le, Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M. and Pyšek, P. (2012) The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.*, **109**, 19–45.

Belser, C., Istace, B., Denis, E., et al. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants*, **4**, 879–887.

Bennetzen, J.L. and Freeling, M. (1993) Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet.*, **9**, 259–261.

Berdan, E.L., Blanckaert, A., Slotte, T., Suh, A., Westram, A.M. and Fragata, I. (2021) Unboxing mutations: Connecting mutation types with evolutionary consequences. *Mol. Ecol.*, **30**, 2710–2723.

Bewick, A.J. and Schmitz, R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.*, **36**, 103–110.

Birchler, J.A. and Veitia, R.A. (2012) Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 14746–14753.

Birchler, J.A. and Veitia, R.A. (2019) Genomic Balance and Speciation. *Epigenetics Insights*, **12**, 1–3.

Birchler, J.A. and Veitia, R.A. (2011) Protein-protein and protein-DNA dosage balance and differential paralog transcription factor retention in polyploids. *Front. Plant Sci.*, **2**, 1–3.

Birchler, J.A. and Veitia, R.A. (2014) The gene balance hypothesis: Dosage effects in plants. *Methods Mol. Biol.*, **1112**, 25–32.

Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D. and Veitia, R.A. (2010) Heterosis. *Plant Cell*, **22**, 2105–2112.

Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.*, **13**, 137–144.

Bomblies, K. (2020) When everything changes at once: Finding a new normal after genome duplication. *Proc. R. Soc. B*, **287**.

Bottani, S., Zabet, N.R., Wendel, J.F. and Veitia, R.A. (2018) Gene Expression Dominance in Allopolyploids: Hypotheses and Models. *Trends Plant Sci.*, **23**, 393–402.

Bourque, G., Pevzner, P.A. and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Res.*, **14**, 507–516.

Branton, D., Deamer, D.W., Marziali, A., et al. (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.

Brenchley, R., Spannagl, M., Pfeifer, M., et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Brochmann, C., Brysting, A.K., Alsos, I.G., Borgen, L., Grundt, H.H., Scheen, A.-C. and Elven, R. (2004) Polyploidy in arctic plants. *Biol. J. Linn. Soc.*, **82**, 521–536.

Brown, T.A., Jones, M.K., Powell, W. and Allaby, R.G. (2009) The complex origins of domesticated crops

in the Fertile Crescent. *Trends Ecol. Evol.*, **24**, 103–109.

Buggs, R.J.A., Chamala, S., Wu, W., Tate, J.A., Schnable, P.S., Soltis, D.E., Soltis, P.S. and Barbazuk, W.B. (2012) Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.*, **22**, 248–252.

Buggs, R.J.A., Elliott, N.M., Zhang, L., Koh, J., Viccini, L.F., Soltis, D.E. and Soltis, P.S. (2010) Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *New Phytol.*, **186**, 175–183.

Bukowski, R., Guo, X., Lu, Y., et al. (2018) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, 1–12.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E. and Baudot, A. (2021) Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.*, **12**, 1–12.

Cao, J., Schneeberger, K., Ossowski, S., et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, **43**, 956–965.

Carlson, K.D., Bhogale, S., Anderson, D., Zaragoza-Mendoza, A. and Madlung, A. (2020) Subfunctionalization of phytochrome B1/B2 leads to differential auxin and photosynthetic responses. *Plant Direct*, **4**, 1–12.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Cavé-Radet, A., Salmon, A., Lima, O., Ainouche, M.L. and Amrani, A. El (2019) Increased tolerance to organic xenobiotics following recent allopolyploidy in Spartina (Poaceae). *Plant Sci.*, **280**, 143–154.

Chalhoub, B., Denoeud, F., Liu, S., et al. (2014) Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. *Science*, **345**, 950–953.

Chang, F., Gu, Y., Ma, H. and Yang, Z. (2013) AtPRK2 promotes ROP1 activation via RopGEFs in the control of polarized pollen tube growth. *Mol. Plant*, **6**, 1187–1201.

Chao, Y., Yuan, J., Li, S., Jia, S., Han, L. and Xu, L. (2018) Analysis of transcripts and splice isoforms in red clover (Trifolium pratense L.) by by single-molecule long-read sequencing. *Plant Mol. Biol.*, **99**, 219–235.

Chapman, B.A., Bowers, J.E., Feltus, F.A. and Paterson, A.H. (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2730–2735.

Charif, D. and Lobry, J.R. (2007) SeqinR 1 . 0-2 : A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis Seqin R 1 . 0-2 : a contributed package to the R project for statistical computing devoted to biological sequences ret. In *Structural approaches to sequence evolution: Molecules, networks, populations*. pp. 207–232.

Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F. and Becker, J. (2020) Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief. Bioinform.*, **21**, 541–552.

Chen, X., Ge, X., Wang, J., Tan, C., King, G.J. and Liu, K. (2015) Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in Brassica rapa suggests that epigenetic modifications play a key role in polyploid genome evolution. *Front. Plant Sci.*, **6**, 1–12.

Chen, Y., Pal, B., Visvader, J.E., Smyth, G.K., Andrews, S. and Macdonald, J.W. (2018) Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR.

F1000Research, 1–40.

Cheng, F., Sun, R., Hou, X., et al. (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nat. Genet.*, **48**, 1218–1224.

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

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G. and Wang, X. (2012) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One*, **7**.

Cheng, X.F., Wittich, P.E., Kieft, H., Angenent, G., XuHan, X. and Lammeren, A.A.M. Van (2000) Temporal and spatial expression of MADS box genes, FBP7 and FBP11, during initiation and early development of ovules in wild type and mutant Petunia hybrida. *Plant Biol.*, **2**, 693–702.

Choulet, F., Alberti, A., Theil, S., et al. (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**, 1249721–1249721.

Chu, E. and Swomley, B. (1961) Chromosomes of lemurine lemurs. *Science*, **133**, 1925–1926.

Chuang, T.J., Chen, F.C. and Chen, Y.Z. (2012) Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 15841–15846.

Chueasiri, C., Chunthong, K., Pitnjam, K., et al. (2014) Rice ORMDL controls sphingolipid homeostasis affecting fertility resulting from abnormal pollen development. *PLoS One*, **9**.

Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P. and Jackson, S.A. (2015) Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Mol. Plant*, **8**, 831–846.

Coate, J.E., Farmer, A.D., Schiefelbein, J.W. and Doyle, J.J. (2020) Expression Partitioning of Duplicate Genes at Single Cell Resolution in Arabidopsis Roots. *Front. Genet.*, **11**.

Cohen, D., Chumakov, I. and Weissenbach, J. (1993) A first-generation physical map of the human genome. *Nature*, **366**, 698–701.

Colombo, L., Franken, J., Krol, A.R. Van Der, Wittich, P.E., Dons, H.J.M. and Angenent, G.C. (1997) Downregulation of ovule-specific MADS box genes from petunia results in maternally controlled defects in seed development. *Plant Cell*, **9**, 703–715.

Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, **6**, 836–846.

Conant, G.C., Birchler, J.A. and Pires, J.C. (2014) Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.*, **19**, 91–98.

Condamine, F.L., Silvestro, D., Koppelhus, E.B. and Antonelli, A. (2020) The rise of angiosperms pushed conifers to decline during global cooling. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 28867–28875.

Conesa, A., Madrigal, P., Tarazona, S., et al. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 1–19.

Crick, F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561–563.

Crick, F. and Watson, J. (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737–738.

Crow, T., Ta, J., Nojoomi, S., Aguilar-Rangel, M.R., Rodríguez, J.V.T., Gates, D., Rellán-Álvarez, R., Sawers, R. and Runcie, D. (2020) Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS Genet.*, **16**.

D'Ario, M. and Sablowski, R. (2019) Cell Size Control in Plants. *Annu. Rev. Genet.*, **53**, 45–65.

D'hont, A., Denoeud, F., Aury, J.M., et al. (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.

Dahm, R. (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, **122**, 565–581.

Darwin, C.R. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* J. Murray, ed., London.

Davis, M.P.A., Dongen, S. Van, Abreu-goodger, C., Bartonicek, N. and Enright, A.J. (2023) Kraken : A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.

Deamer, D., Akeson, M. and Branton, D. (2016) Three decades of nanopore sequencing. *Nat. Biotechnol.*, **34**, 518–524.

Dean, E.J., Davis, J.C., Davis, R.W. and Petrov, D.A. (2008) Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.*, **4**.

Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**.

DeSmet, R., Adams, K.L., Vandepoele, K., Montagu, M.C.E. Van, Maere, S. and Peer, Y. Van de (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci.*, **110**, 2898–2903.

Devos, K.M. and Gale, M.D. (2000) Genome relationships: The grass model in current research. *Plant Cell*, **12**, 637–646.

Diez, C.M., Roessler, K. and Gaut, B.S. (2014) Epigenetics and plant genome evolution. *Curr. Opin. Plant Biol.*, **18**, 1–8.

Donato, A. Di, Filippone, E., Ercolano, M.R. and Frusciante, L. (2018) Genome Sequencing of Ancient Plant Remains: Findings, Uses and Potential Applications for the Study and Improvement of Modern Crops. *Front. Plant Sci.*, **9**.

Donis-Keller, H., Green, P., Helms, C., et al. (1987) A genetic linkage map of the human genome. *Cell*, **51**, 319–337.

Doyle, J.J. and Coate, J.E. (2019) Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.*, **180**, 1–52.

Drake, P.L., Froend, R.H. and Franks, P.J. (2013) Smaller, faster stomata: Scaling of stomatal size, rate of response, and stomatal conductance. *J. Exp. Bot.*, **64**, 495–505.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and DePamphilis, C.W. (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol. Biol. Evol.*, **23**, 469–478.

Dubcovsky, J., Luo, M.C. and Dvořák, J. (1995) Differentiation between homoeologous chromosomes 1A of wheat and 1Am of Triticum monococcum and its recognition by the wheat Ph1 locus. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 6645–6649.

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C.M., Deal, K.R., Dai, X. and Dawson, M.W. (2018) Structural variation and rates of genome evolution in the grass family seen through comparison of sequences of genomes greatly differing in size. *Plant J.*, **95**, 487–503.

Dykhuizen, D. and Hartl, D. (1981) Evolution of Competitive Ability in Escherichia coli. *Evolution (N. Y).*,

35, 581.

Edgar, R.C. (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 1–19.

Edger, P.P., Poorten, T.J., Vanburen, R., et al. (2019) Origin and evolution of the octoploid strawberry genome. *Nat. Genet.*, **51**, 541–547.

Edger, P.P., Smith, R., McKain, M.R., et al. (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell*, **29**, 2150–2167.

Eid, J., Fehr, A., Gray, J., et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

El-Sharkawy, I., Mila, I., Bouzayen, M. and Jayasankar, S. (2010) Regulation of two germin-like protein genes during plum fruit development. *J. Exp. Bot.*, **61**, 1761–1770.

Emerson, R.A., Beadle, G.W. and Fraser, A.C. (1935) *A Summary of Linkage Studies in Maize Cornell University*, Cornell University Agricultural Experiment Station.

Ernest, L., Lam, E.T., Hastie, A., et al. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.*, **30**, 771–776.

Escudero, M. and Wendel, J.F. (2020) The grand sweep of chromosomal evolution in angiosperms. *New Phytol.*, **228**, 805–808.

Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G., Yanofsky, M.F., Kater, M.M. and Colombo, L. (2003) MADS-Box Protein Complexes Control Carpel and Ovule Development in Arabidopsis. *Plant Cell*, **15**, 2603–2611.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Fleischmann, R.D., Adams, M.D., White, O., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

Flusberg, B.A., Webster, D., Lee, J., Travers, K., Olivares, E., Clark, A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.

Folta, K.M. and Barbey, C.R. (2019) The strawberry genome: a complicated past and promising future. *Hortic. Res.*, **6**, 19–21.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–1545.

Forrester, N.J., Rebolleda-Gómez, M., Sachs, J.L. and Ashman, T.L. (2020) Polyploid plants obtain greater fitness benefits from a nutrient acquisition mutualism. *New Phytol.*, **227**, 944–954.

Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L. and Peer, Y. Van de (2020) Polyploidy: A Biological Force From Cells to Ecosystems. *Trends Cell Biol.*, **30**, 688–694.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.*, **60**, 433–453.

Freeling, M. (2017) Picking up the ball at the K/Pg boundary: The distribution of ancient polyploidies in the plant phylogenetic tree as a spandrel of asexuality with occasional sex. *Plant Cell*, **29**, 202–206.

Freeling, M., Scanlon, M.J. and Fowler, J.F. (2015) Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.*, **35**, 110–118.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D. and Schnable, J.C. (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr. Opin. Plant Biol.*, **15**, 131–139.

Gabur, I., Singh, H., Rod, C. and Isobel, J.S. (2019) Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.*, **132**, 733–750.

Gale, M.D. and Devos, K.M. (1998) Plant comparative genetics after 10 years. *Science*, **282**, 656–659.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A. and Freeling, M. (2014) Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.*, **31**, 448–454.

Gebhardt, C., Ritter, E., Barone, A., et al. (1991) RFLP maps of potato and their alignment with the homoeologous tomato genome. *Theor. Appl. Genet.*, **83**, 49–57.

Gepts, P. (2004) Crop Domestication as a Long-term Selection Expermient. *Plant Breed. Rev.*, **24**, 1–442.

Gerstein, A.C., Chun, H.J.E., Grant, A. and Otto, S.P. (2006) Genomic convergence toward diploidy in Saccharomyces cerevisiae. *PLoS Genet.*, **2**, 1396–1401.

Gerstein, A.C., Lim, H., Berman, J. and Hickman, M.A. (2017) Ploidy tug-of-war: Evolutionary and genetic environments influence the rate of ploidy drive in a human fungal pathogen. *Evolution (N. Y).*, **71**, 1025–1038.

Gill, B.S. and Kimber, G. (1974) Giemsa C banding and the evolution of wheat. *Proc. Natl. Acad. Sci. U. S. A.*, **71**, 4086–4090.

Glastad, K.M., Gokhale, K., Liebig, J. and Goodisman, M.A.D. (2016) The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Sci. Rep.*, **6**, 1–14.

Glémin, S., Scornavacca, C., Dainat, J., et al. (2019) Pervasive hybridizations in the history of wheat relatives. *Sci. Adv.*, **5**, 1–10.

Glover, N.M., Redestig, H. and Dessimoz, C. (2016) Homoeologs: What Are They and How Do We Infer Them? *Trends Plant Sci.*, **21**, 609–621.

Goidts, V., Szamalek, J.M., Jong, P.J. De, Cooper, D.N., Chuzhanova, N., Hameister, H. and Kehrer-Sawatzki, H. (2005) Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res.*, **15**, 1232–1242.

Golicz, A.A., Batley, J. and Edwards, D. (2016) Towards plant pangenomics. *Plant Biotechnol. J.*, **14**, 1099–1105.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., et al. (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.*, **8**.

Gout, J.F. and Lynch, M. (2015) Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol. Biol. Evol.*, **32**, 2141–2148.

Grandont, L., Jenczewski, E. and Lloyd, A. (2013) Meiosis and its deviations in polyploid plants. *Cytogenet. Genome Res.*, **140**, 171–184.

Green, R.E., Krause, J., Briggs, A.W., et al. (2010) A draft sequence of the neandertal genome. *Science*, **328**, 710–722.

Gu, Y.Q., Anderson, O.D., Londeorë, C.F., Kong, X., Chibbar, R.N. and Lazo, G.R. (2003) Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat

genomes. *Genome*, **46**, 1084–1097.

Gunn, B.F., Murphy, D.J., Walsh, N.G., Conran, J.G., Pires, J.C., Macfarlane, T.D. and Birch, J.L. (2020) Evolution of Lomandroideae: Multiple origins of polyploidy and biome occupancy in Australia. *Mol. Phylogenet. Evol.*, **149**, 106836.

Guo, M., Davis, D. and Birchler, J.A. (1996) Dosage effects on gene expression in a maize ploidy series. *Genetics*, **142**, 1349–1355.

Hahn, M.A., Kleunen, M. van and Müller-Schärer, H. (2012) Increased Phenotypic Plasticity to Climate May Have Boosted the Invasion Success of Polyploid Centaurea stoebe. *PLoS One*, **7**.

Hancock, J.F. (2005) Contributions of domesticated plant studies to our understanding of plant evolution. *Ann. Bot.*, **96**, 953–963.

Hannweg, K., Steyn, W. and Bertling, I. (2016) In vitro-induced tetraploids of Plectranthus esculentus are nematode-tolerant and have enhanced nutritional value. *Euphytica*, **207**, 343–351.

Hao, Yue, Mabry, M.E., Edger, P.P., et al. (2021) The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. *Genome Res.*, **31**, 799–810.

Hao, Yaqi, Zong, X., Ren, P., Qian, Y. and Fu, A. (2021) Basic helix-loop-helix (Bhlh) transcription factors regulate a wide range of functions in arabidopsis. *Int. J. Mol. Sci.*, **22**.

Herendeen, P.S., Friis, E.M., Pedersen, K.R. and Crane, P.R. (2017) Palaeobotanical redux: Revisiting the age of the angiosperms. *Nat. Plants*, **3**, 1–8.

Hias, N., Svara, A. and Keulemans, J.W. (2018) Effect of polyploidisation on the response of apple (Malus × domestica Borkh.) to Venturia inaequalis infection. *Eur. J. Plant Pathol.*, **151**, 515–526.

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R. and Leitch, I.J. (2017) Is There an Upper Limit to Genome Size? *Trends Plant Sci.*, **22**, 567–573.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A. and Haeseler, A. Von (2018) MPBoot : fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.*, **18**, 1–11.

Hoffmann, R.D. and Palmgren, M. (2016) Purifying selection acts on coding and non-coding sequences of paralogous genes in Arabidopsis thaliana. *BMC Genomics*, **17**, 1–13.

Hollister, J.D. (2015) Polyploidy: Adaptation to the genomic environment. *New Phytol.*, **205**, 1034–1039.

Hsu, T.C. and Arrighi, F.E. (1971) Distribution of constitutive heterochromatin in mammalian chromosomes. *Chromosoma*, **34**, 243–253.

Hu, G. and Wendel, J.F. (2019) Cis–trans controls and regulatory novelty accompanying allopolyploidization. *New Phytol.*, **221**, 1691–1700.

Hu, Q., Ma, Y., Mándaková, T., et al. (2021) Genome evolution of the psammophyte Pugionium for desert adaptation and further speciation. *Proc. Natl. Acad. Sci. U. S. A.*, **118**.

Huang, X., Lu, T. and Han, B. (2013) Resequencing rice genomes: An emerging new era of rice genomics. *Trends Genet.*, **29**, 225–232.

Hufnagel, B., Marques, A., Soriano, A., et al. (2020) High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nat. Commun.*, **11**, 1–12.

Hughes, T.E., Langdale, J.A. and Kelly, S. (2014) The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.*, **24**, 1348–1355.

Iannicelli, J., Guariniello, J., Tossi, V.E., Regalado, J.J., Ciaccio, L. Di, Baren, C.M. van, Pitta Álvarez, S.I.

and Escandón, A.S. (2020) The “polyploid effect” in the breeding of aromatic and medicinal species. *Sci. Hortic. (Amsterdam).*, **260**, 108854.

Ilyas, M., Rasheed, A. and Mahmood, T. (2016) Functional characterization of germin and germin-like protein genes in various plant species using transgenic approaches. *Biotechnol. Lett.*, **38**, 1405–1421.

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

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

IRGSP (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Jaillon, O., Aury, J.M., Noel, B., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jan, H.U., Guan, M., Yao, M., et al. (2019) Genome-wide haplotype analysis improves trait predictions in Brassica napus hybrids. *Plant Sci.*, **283**, 157–164.

Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., et al. (2017) The genome of Chenopodium quinoa. *Nature*, **542**, 307–312.

Jenczewski, E. and Alix, K. (2004) From Diploids to Allopolyploids: The Emergence of Efficient Pairing Control Genes in Plants. *CRC. Crit. Rev. Plant Sci.*, **23**, 21–45.

Jenuwein, T. and Allis, C.D. (2001) Translating the Histone Code. *Science*, **293**, 1074–1080.

Jia, J., Zhao, S., Kong, X., et al. (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.

Jiao, W.B., Accinelli, G.G., Hartwig, B., et al. (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.*, **27**, 778–786.

Jiao, W.B. and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.*, **36**, 64–70.

Jiao, Y., Li, J., Tang, H. and Paterson, A.H. (2014) Integrated Syntenic and Phylogenomic Analyses Reveal an Ancient Genome Duplication in Monocots. *Plant Cell*, **26**, 2792–2802.

Joron, M., Frezal, L., Jones, R.T., et al. (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Juery, C., Concia, L., Oliveira, R. De, Papon, N. and Ramírez-, R. (2020) New insights into homoeologous copy number variations in the hexaploid wheat genome. *bioRxiv*, 1–43.

Jung, H., Ventura, T., Sook Chung, J., Kim, W.J., Nam, B.H., Kong, H.J., Kim, Y.O., Jeon, M.S. and Eyun, S. Il (2020) Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput. Biol.*, **16**, 1–25.

Kashkush, K., Feldman, M. and Levy, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.*, **33**, 102–106.

Keller, B. and Feuillet, C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.*, **5**, 246–251.

Kellis, M., Birren, B.W. and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome

duplication in the yeast. *Saccharomyces cerevisiae. Nat.*, **428:617–62**, 617–624.

Kersey, P.J. (2019) Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.*, **48**, 1–8.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**.

Kim, K. Do, Baidouri, M. El, Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grimwood, J. and Jackson, S.A. (2015) A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant Physiol.*, **168**, 1433–1447.

Kohara, Y., Akiyama, K. and Isono, K. (1987) The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, **50**, 495–508.

Koide, Y., Kuniyoshi, D. and Kishima, Y. (2020) Fertile Tetraploids: New Resources for Future Rice Breeding? *Front. Plant Sci.*, **11**, 1–6.

Kovach, M.J., Sweeney, M.T. and McCouch, S.R. (2007) New insights into the history of rice domestication. *Trends Genet.*, **23**, 578–587.

Krueger, F. and Andrews, S.R. (2011) Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications. , **27**, 1571–1572.

Lacoste, C., Fabre, A., Pécheux, C., Lévy, N., Krahn, M., Malzac, P., Bonello-Palot, N., Badens, C. and Bourgeois, P. (2017) Le séquençage d'ADN à haut débit en pratique clinique. *Arch. Pediatr.*, **24**, 373–383.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C. and Soltis, P.S. (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.*, **105**, 348–363.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004) Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics*, **166**, 935–945.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. , **10**.

Law, J.A. and Jacobsen, S.E. (2011) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet.*, **11**, 204–220.

LeComber, S. and Smith, C. (2004) Polyploidy in fishes: patterns and processes. *Biol. J. Linn. Soc.*, **82**, 431–442.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., et al. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.

Leitch, A.R. and Leitch, I.J. (2008) Genomic plasticity and the diversity of polyploid plants. *Science*, **320**, 481–483.

Leitch, I.J. and Bennett, M.D. (1997) Polyploidy in angiosperms. *Trends Plant Sci.*, **2**, 470–476.

Levene, H.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. and Webb, W.W. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.

Levin, D.A. and Soltis, D.E. (2018) ScienceDirect Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K – Pg interlude. *Curr. Opin. Plant Biol.*, **42**, 1–7.

Li, A., Geng, S., Zhang, L., Liu, D. and Mao, L. (2015) Making the Bread : Insights from Newly Synthesized Allohexaploid Wheat. , 847–859.

Li, B. and Dewey, C.N. (2011) RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 1–16.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics*, **00**, 1–3.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M. and Muehlbauer, G.J. (2016) Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC Genomics*, **17**, 875.

Li, M., Wang, R., Wu, X. and Wang, J. (2020) Homoeolog expression bias and expression level dominance (ELD) in four tissues of natural allotetraploid Brassica napus. *BMC Genomics*, **21**, 1–15.

Li, N., Xu, C., Zhang, A., Lv, R., Meng, X., Lin, X., Gong, L., Wendel, J.F. and Liu, B. (2019) DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. *New Phytol.*, **223**, 979–992.

Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K., Li, L. and Zhang, S. (2019) Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). *Hortic. Res.*, **6**, 1–12.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H. and Barker, M.S. (2015) Early genome duplications in conifers and other seed plants. *Sci. Adv.*, **1**, 1–7.

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J. and Barker, M.S. (2018) Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 4713–4718.

Lian, S., Zhou, Y., Liu, Z., Gong, A. and Cheng, L. (2020) The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences. *BMC Plant Biol.*, **20**, 1–16.

Liao, B.Y. and Zhang, J. (2006) Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol. Biol. Evol.*, **23**, 1119–1128.

Lieberman-aiden, E., Berkum, N.L. Van, Williams, L., et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–294.

Lightfoot, D.J., Jarvis, D.E., Ramaraj, T., Lee, R., Jellen, E.N. and Maughan, P.J. (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. *BMC Biol.*, **15**, 1–15.

Lim, C.W., Baek, W., Han, S.W. and Lee, S.C. (2013) Arabidopsis PYl8 plays an important role for ABA signaling and drought stress responses. *Plant Pathol. J.*, **29**, 471–476.

Ling, H.Q., Wang, Jun, Zhao, S., et al. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, **496**, 87–90.

Lisch, D. (2009) Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.*, **60**, 43–66.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**.

Liu, S., Liu, Y., Yang, X., et al. (2014) The brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.*, **5**, 1–11.

Liu, X.L., Jiang, F.F., Wang, Z.W., et al. (2017) Wider geographic distribution and higher diversity of hexaploids than tetraploids in Carassius species complex reveal recurrent polyploidy effects on adaptive evolution. *Sci. Rep.*, **7**, 1–10.

Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N., Sun, S., Zhang, L. and Wang, X. (2017) Two highly similar poplar paleo-subgenomes suggest an autotetraploid ancestor of salicaceae plants. *Front. Plant Sci.*, **8**, 1–11.

Liu, Z.-J.J., Cai, jing, Peer, Y. Van de and Liu, Z.-J.J. (2015) The genome sequence of the orchid Phalaenopsis equestris. *Nat. Genet.*, **47**, 65–72.

Lloyd, A. and Bomblies, K. (2016) Meiosis in autopolyploid and allopolyploid Arabidopsis. *Curr. Opin. Plant Biol.*, **30**, 116–122.

Lockhart D. J., Dong, H., Byrne, M.C., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, **21**, 597–614.

Lohaus, R. and Peer, Y. Van de (2016) Of dups and dinos: Evolution at the K/Pg boundary. *Curr. Opin. Plant Biol.*, **30**, 62–69.

Lopes, F.R., Jjingo, D., Silva, C.R.M. Da, et al. (2013) Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. *PLoS One*, **8**.

Lovell, J.T., MacQueen, A.H., Mamidi, S., et al. (2021) Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, **590**, 438–444.

Lu, K., Wei, L., Li, X., et al. (2019) Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat. Commun.*, **10**, 1–12.

Luo, M.-C., Gu, Y.Q., You, F.M., et al. (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. *Proc. Natl. Acad. Sci.*, **110**, 7940–7945.

Luo, Z., Iaffaldano, B.J. and Cornish, K. (2018) Colchicine-induced polyploidy has the potential to improve rubber yield in Taraxacum kok-saghyz. *Ind. Crops Prod.*, **112**, 75–81.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Ma, Y.Q., Pu, Z.Q., Tan, X.M., Meng, Q., Zhang, K.L., Yang, L., Ma, Y.Y., Huang, X. and Xu, Z.Q. (2022) SEPALLATA-like genes of Isatis indigotica can affect the architecture of the inflorescences and the development of the floral organs. *PeerJ*, **10:e13034**, 1–25.

Maccaferri, M., Harris, N., Twardziok, S.O., et al. (2019) Durum wheat genome reveals past domestication signatures and future improvement targets. *Nat. Genet.*, **In Press**.

Maddox, B. (2003) The double helix and the "wronged heroine." *Nature*, **421**, 407–408.

Madlung, A., Tyagi, A.P., Watson, B., Jiang, H., Kagochi, T., Doerge, R.W., Martienssen, R. and Comai, L. (2005) Genomic changes in synthetic Arabidopsis polyploids. *Plant J.*, **41**, 221–230.

Mandáková, T., Li, Z., Barker, M.S. and Lysak, M.A. (2017) Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant J.*, **91**, 3–21.

Mandáková, T. and Lysak, M.A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.*, **42**, 55–65.

Marais, D.L. Des and Rausher, M.D. (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, **454**, 762–765.

Margulies, M., Egholm, M., Altman, W.E., et al. (2005) Genome sequencing in microfabricated high-

density picolitre reactors. *Nature*, **437**, 376–380.

Maria Maggiolini, F.A., Sanders, A.D., Shew, C.J., et al. (2020) Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.*, **30**, 1680–1693.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 10–12.

Mascher, M., Wicker, T., Jenkins, J., et al. (2021) Long-read sequence assembly: a technical evaluation in barley. *Plant Cell*, **33**, 1888–1906.

Mason, A.S. and Pires, J.C. (2015) Unreduced gametes: Meiotic mishap or evolutionary mechanism? *Trends Genet.*, **31**, 5–10.

Mason, A.S. and Wendel, J.F. (2020) Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. *Front. Genet.*, **11**, 1–10.

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Krüger, D.H. and Terauchi, R. (2005) SuperSAGE. *Cell. Microbiol.*, **7**, 11–18.

Mattingly, K.Z. and Hovick, S.M. (2021) Autopolyploids of Arabidopsis are more phenotypically plastic than their diploid progenitors. *Ann. Bot.*, 1–13.

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *PNAS*, **74**, 560–564.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Magnuson-Ford, K., Barker, M.S., Rieseberg, L.H. and Otto, S.P. (2011) Recently formed polyploid plants diversify at lower rates. *Science*, **333**, 1257.

Mazoyer, M. and Roudart, L. (1997) *Histoire des agricultures du monde. Du néolithique à la crise contemporaine* Editions d.,.

Mccarthy, D.J., Chen, Y. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

McClintock, B. (1929) A Cytological and Genetical Study of Triploid Maize. *Genetics*, **14**, 180–222.

Meeus, S., Semberova, K., Storme, N. De, Geelen, D. and Vallejo-marı, M. (2020) Effect of Whole-Genome Duplication on the Evolutionary Rescue of Sterile Hybrid Monkeyflowers.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M. (2015) The molecular biology of meiosis in plants. *Annu. Rev. Plant Biol.*, **66**, 297–327.

Meyer, A. and Schartl, M. (1999) Gene and genome duplications in vertebrates: The one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.*, **11**, 699–704.

Michael, T.P. and Jackson, S. (2013) The First 50 Plant Genomes. *Plant Genome*, **6**, plantgenome2013.03.0001in.

Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D. and Duret, L. (2012) High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics*, **28**, 1078–1085.

Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.

Mitra, R.D., Shendure, J., Olejnik, J. and Church, G.M. (2003) Fluorescent in situ sequencing on polymerase colonies. , **320**, 55–65.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., et al. (2022) Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, **602**, 101–105.

Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995) Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.

Moore, G., Roberts, M., Aragon-Alcaide, L. and Foote, T. (1997) Centromeric sites and cereal chromosome evolution. *Chromosoma*, **105**, 321–323.

Mouse genome sequencing consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

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

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. and Salse, J. (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.*, **16**, 262.

Murat, F., Zhang, R., Guizard, S., et al. (2014) Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biol. Evol.*, **6**, 12–33.

Murat, F., Zhang, R., Guizard, S., Gavranovic, H., Flores, R., Steinbach, D., Quesneville, H., Tannier, E. and Salse, J. (2015) Karyotype and Gene Order Evolution from Reconstructed Extinct Ancestors Highlight Contrasts in Genome Plasticity of Modern Rosid Crops. *Genome Biol. Evol.*, **7**, 735–749.

Murphy, W.J., Larkin, D.M., Everts-Van Der Wind, A., et al. (2005) Evolution: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**, 613–617.

Mutti, J.S., Bhullar, R.K. and Gill, K.S. (2017) Evolution of gene expression balance among homeologs of natural polyploids. *G3 Genes, Genomes, Genet.*, **7**, 1225–1237.

Myers, E.W., Sutton, G.G., Delcher, A., et al. (2000) A Whole-Genome Assembly of Drosophila. *Science*, **287**, 2196–2204.

Nakatani, Y., Takeda, H., Kohara, Y. and Morishita, S. (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, **17**, 1254–1265.

Naser-Khdour, S., Quang Minh, B., Zhang, W., Stone, E.A. and Lanfear, R. (2019) The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biol. Evol.*, **11**, 3341–3352.

Nguyen, L., Schmidt, H.A., Haeseler, A. Von and Minh, B.Q. (2014) IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

Ohno, Susumu (1970) *Evolution by Gene Duplication*, Springer, Berlin, Heidelberg.

Ohno, S (1970) *Evolution by Gene Duplication*, Springer-Verlag.

Olmo, E. (1983) Nucleotype and cell size in vertebrates: a review. *basic Appl. Histochem.*, **27**, 227–256.

Oswald, B.P. and Nuismer, S.L. (2007) Neopolyploidy and pathogen resistance. *Proc. R. Soc. B Biol. Sci.*, **274**, 2393–2397.

Page, J.T., Liechty, Z.S., Alexander, R.H., Clemons, K., Hulse-Kemp, A.M., Ashrafi, H., Deynze, A. Van, Stelly, D.M. and Udall, J.A. (2016) DNA Sequence Evolution and Rare Homoeologous Conversion in Tetraploid Cotton. *PLoS Genet.*, **12**, 1–22.

Palmer, S.A., Clapham, A.J., Rose, P., Freitas, F.O., Owen, B.D., Beresford-Jones, D., Moore, J.D., Kitchen, J.L. and Allaby, R.G. (2012) Archaeogenomic evidence of punctuated genome evolution in gossypium. *Mol. Biol. Evol.*, **29**, 2031–2038.

Panchy, N., Lehti-Shiu, M. and Shiu, S.H. (2016) Evolution of gene duplication in plants. *Plant Physiol.*, **171**, 2294–2316.

Pandit, M.K., Pocock, M.J.O. and Kunin, W.E. (2011) Ploidy influences rarity and invasiveness in plants. *J. Ecol.*, **99**, 1108–1115.

Pankin, A., Altmüller, J., Becker, C. and Korff, M. von (2018) Targeted resequencing reveals genomic signatures of barley domestication. *New Phytol.*, **218**, 1247–1259.

Parkin, I.A.P., Koh, C., Tang, H., et al. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol.*, **15**, 1–18.

Parks, M.B., Nakov, T., Ruck, E.C., Wickett, N.J. and Alverson, A.J. (2018) Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *Am. J. Bot.*, **105**, 330–347.

Passarge, E., Horsthemke, B. and Farber, R.A. (1999) Incorrect use of the term synteny. *Nat. Genet.*, **23**, 387.

Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Paterson, A.H., Wendel, J.F., Gundlach, H., et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, **492**, 423–427.

Payne, A., Holmes, N., Rakyan, V. and Loose, M. (2018) Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*, 1–27.

Pécrix, Y., Rallo, G., Folzer, H., Cigna, M., Gudin, S. and Bris, M. Le (2011) Polyploidization mechanisms: Temperature environment can induce diploid gamete formation in Rosa sp. *J. Exp. Bot.*, **62**, 3587–3597.

Peer, Y. Van de, Ashman, T.L., Soltis, P.S. and Soltis, D.E. (2021) Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell*, **33**, 11–26.

Peer, Y. Van De, Mizrachi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.*, **18**, 411–424.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 9748–9753.

Pham, S.K. and Pevzner, P.A. (2010) DRIMM-Synteny: Decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.

Piperno, D.R., Ranere, A.J., Holst, I., Iriarte, J. and Dickau, R. (2009) Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 5019–5024.

Poggio, L. and González, G.E. (2018) Cytological diploidization of paleopolyploid genus Zea: Divergence between homoeologous chromosomes or activity of pairing regulator genes? *PLoS One*, **13**, 1–17.

Poinar, H.N., Schwarz, C., Qi, J., et al. (2006) Metagenomics to Paleogenomics : *Science*, **311**, 392–395.

Pons, P. and Latapy, M. (2006) Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, **10**, 191–218.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-korts, D. and Goué, N. (2019) Tracing the ancestry of modern bread wheats. *Nat. Genet.*, **51**, 905–911.

Pont, C., Murat, F., Guizard, S., et al. (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant J.*, **76**, 1030–1044.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C. and Salse, J. (2019) Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.*, **20**, 1–17.

Price, H.J. (1988) DNA Content Variation among Higher Plants. *Ann. Missouri Bot. Gard.*, **75**, 1248–1257.

Purugganan, M.D. and Jackson, S.A. (2021) Advancing crop genomics from lab to field. *Nat. Genet.*, **53**,

595–601.

Purushothaman, S., Toumazou, C. and Ou, C.P. (2006) Protons and single nucleotide polymorphism detection: A simple use for the Ion Sensitive Field Effect Transistor. *Sensors Actuators, B Chem.*, **114**, 964–968.

Putnam, N.H., Connell, B.O., Stites, J.C., Rice, B.J., Hartley, P.D., Sugnet, C.W., Haussler, D. and Rokhsar, D.S. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.*, **26**, 342–350.

Qian, W., Liao, B.-Y., Chang, A.Y.F. and Zhang, J. (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.*, **26**, 425–430.

Ramírez-González, R.H., Borrill, P., Lang, D., et al. (2018) The transcriptional landscape of polyploid wheat. *Science*, **361**, 1–12.

Rang, F.J., Kloosterman, W.P. and Ridder, J. de (2018) From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.*, **19**, 1–11.

Rapp, R.A., Udall, J.A. and Wendel, J.F. (2009) Genomic expression dominance in allopolyploids. *BMC Biol.*, **7**, 1–10.

Raymond, O., Gouzy, J., Just, J., et al. (2018) The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.*, **50**, 772–777.

Rejlová, L., Chrtek, J., Trávníček, P., Lučanová, M., Vít, P. and Urfus, T. (2019) Polyploid evolution: The ultimate way to grasp the nettle. *PLoS One*, **14**, 1–24.

Renne, P.R., Sprain, C.J., Richards, M.A., Self, S., Vanderkluysen and L., Pande, K. (2015) State shift in Deccan volcanism at the Cretaceous-Paleogene boundary, possibly induced by impact. *Science*, **350**, 76–78.

Renny-Byfield, S., Gallagher, J.P., Grover, C.E., Szadkowski, E., Page, J.T., Udall, J.A., Wang, X., Paterson, A.H. and Wendel, J.F. (2014) Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.*, **6**, 559–571.

Renny-Byfield, S., Gong, L., Gallagher, J.P. and Wendel, J.F. (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol. Biol. Evol.*, **32**, 1063–1071.

Renny-Byfield, S., Rodgers-Melnick, E. and Ross-Ibarra, J. (2017) Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol.*, **34**, 1825–1832.

Renny-Byfield, S. and Wendel, J.F. (2014) Doubling down on genomes: Polyploidy and crop plants. *Am. J. Bot.*, **101**, 1711–1725.

Renwick, J.H. (1971) The mapping of human chromosomes. *Annu. Rev. Genet.*, **5**, 81–120.

Rho, I.R., Hwang, Y.J., Lee, H. Il, Lim, K.B. and Lee, C.H. (2012) Interspecific hybridization of diploids and octoploids in strawberry. *Sci. Hortic. (Amsterdam).*, **134**, 46–52.

Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., Meiri, S., Belmaker, J. and Mayrose, I. (2019) The global biogeography of polyploid plants. *Nat. Ecol. Evol.*, **3**, 265–273.

Riley, R. and Chapman, victor (1967) Effect of 5BS in suppressing the Expression of Altered Dosage of (5BL on Meiotic Chromosome Pairing in Triticum aestivum. *Nature*, **216**, 60–62.

Riley, R. and Chapman, V. (1958) Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature*, 713–715.

Robinson, M.D., Mccarthy, D.J. and Smyth, G.K. (2010) edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rohart, F., Gautier, B., Singh, A. and Lê Cao, K.A. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.*, **13**, 1–19.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.

Rosellini, D., Ferradini, N., Allegrucci, S., et al. (2016) Sexual polyploidization in Medicago sativa L.: Impact on the phenotype, gene transcription, and genome methylation. *G3 Genes, Genomes, Genet.*, **6**, 925–938.

Ruiz, M., Oustric, J., Santini, J. and Morillon, R. (2020) Synthetic Polyploidy in Grafted Crops. *Front. Plant Sci.*, **11**.

Šafář, J., Bartoš, J., Janda, J., et al. (2004) Dissecting large and complex genomes: Flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.*, **39**.

Salman-Minkov, A., Sabath, N. and Mayrose, I. (2016) Whole-genome duplication as a key factor in crop domestication. *Nat. Plants*, **2**, 1–4.

Salse, J. (2016) Deciphering the evolutionary interplay between subgenomes following polyploidy: A paleogenomics approach in grasses. *Am. J. Bot.*, **103**, 1167–1174.

Salse, J. (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.*, **15**, 122–130.

Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. and Feuillet, C. (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.*, **10**, 619–630.

Salse, J., Bolot, S., Throude, M., et al. (2008) Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. *Plant Cell Online*, **20**, 11–24.

Sanger, F. (2001) The early days of DNA sequences. *Nat. Med.*, **7**, 267–268.

Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, C. and Petersen, G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.*, **162**, 729–773.

Sarda, S., Zeng, J., Hunt, B.G. and Yi, S. V. (2012) The evolution of invertebrate gene body methylation. *Mol. Biol. Evol.*, **29**, 1907–1916.

Sauquet, H., Balthazar, M. Von, Magallón, S., et al. (2017) The ancestral flower of angiosperms and its early diversification. *Nat. Commun.*, **8**.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schmid, M., Evans, B.J. and Bogart, J.P. (2015) Polyploidy in Amphibia. *Cytogenet. Genome Res.*, **145**, 315–330.

Schmutz, J., Cannon, S.B., Schlueter, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.*, **4**, 265–277.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.*, **108**,

4069–4074.

Schubert, I. and Vu, G.T.H. (2016) Genome Stability and Evolution: Attempting a Holistic View. *Trends Plant Sci.*, **21**, 749–757.

Schwartz, D.C., Li, X., Hernandez, L.I., et al. (1993) Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. *Science*, **262**, 110–114.

Scornavacca, C., Jacox, E. and Szöllosi, G.J. (2015) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31**, 841–848.

Selvaraj, S., Dixon, J.R., Bansal, V. and Ren, B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–1118.

Seoighe, C. and Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.*, **2**, 548–554.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. (2017) DNA sequencing at 40: Past, present and future. *Nature*, **550**.

Shendure, J., Porreca, G.J., Reppas, N.B., et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Shi, J., Ma, X., Zhang, J., et al. (2019) Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.*, **10**, 1–9.

Shi, T., Rahmani, R.S., Gugger, P.F., et al. (2020) Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants. *Mol. Biol. Evol.*, **37**, 2394–2413.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 8794–8797.

Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.*, **6**, 715–722.

Slane, D., Reichardt, I., Kasmi, F. El, Bayer, M. and Jürgens, G. (2017) Evolutionarily diverse SYP1 Qa-SNAREs jointly sustain pollen tube growth in Arabidopsis. *Plant J.*, **92**, 375–385.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H. and Hood, L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.

Soltis, D.E., Gitzendanner, M.A., Stull, G., et al. (2013) The potential of genomics in plant systematics. *Taxon*, **62**, 886–898.

Soltis, D.E., Segovia-Salcedo, M.C., Jordon-Thaden, I., et al. (2014) Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytol.*, **202**, 1105–1117.

Soltis, D.E., Visger, C.J., Blaine Marchant, D. and Soltis, P.S. (2016) Polyploidy: Pitfalls and paths to a paradigm. *Am. J. Bot.*, **103**, 1146–1166.

Soltis, P.S. and Soltis, D. (2014) Flower Diversity and Angiosperm Diversifi cation(Flower development: open questions and future directionschapter 4）. *Methods Mol. Biol.*, **1110**, 85–100.

Soltis, P.S. and Soltis, D.E. (2009) The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.*, **60**, 561–588.

Song, K., Li, L. and Zhang, G. (2018) Relationship Among Intron Length, Gene Expression, and Nucleotide Diversity in the Pacific Oyster Crassostrea gigas. *Mar. Biotechnol.*, **20**, 676–684.

Staden, R. (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Res.*, **10**, 4731–4751.

Stamatakis, A. (2014) RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stebbins, G.L. (1947) Types of Polyploids: Their Classification and Significance. *Adv. Genet.*, **1**, 403–429.

Steijger, T., Abril, J.F., Engström, P.G., et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S. V., Obermeier, C., Parkin, I.A.P., Chèvre, A.M. and Snowdon, R.J. (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus. *Plant Biotechnol. J.*, **15**, 1478–1489.

Storme, N. de, Copenhaver, G.P. and Geelen, D. (2012) Production of diploid male gametes in Arabidopsis by cold-induced Destabilization of postmeiotic radial microtubule arrays. *Plant Physiol.*, **160**, 1808–1826.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., et al. (2007) Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics*, **176**, 2055–2067.

Sturtevant, A.H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool.*, **14**, 43–59.

Sun, H., Wu, S., Zhang, G., et al. (2017) Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Mol. Plant*, **10**, 1293–1306.

Sun, Y., Zhong, M., Li, Y., Zhang, R., Su, L., Xia, G. and Wang, H. (2021) GhADF6-mediated actin reorganization is associated with defence against Verticillium dahliae infection in cotton. *Mol. Plant Pathol.*, **22**, 1656–1667.

Suyama, M., Torrents, D., Bork, P. and Delbru, M. (2006) PAL2NAL : robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, 609–612.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

Szöllosi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E. and Daubin, V. (2013) Efficient exploration of the space of reconciled gene trees. *Syst. Biol.*, **62**, 901–912.

Takuno, S. and Gaut, B.S. (2012) Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. *Mol. Biol. Evol.*, **29**, 219–227.

Takuno, S. and Gaut, B.S. (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc. Natl. Acad. Sci.*, **110**, 1797–1802.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A. and Peer, Y. Van de (2003) Genome duplication, a trait shared by 22 000 species of ray-finned fish. *Genome Res.*, **13**, 382–390.

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*, **282**, 2012–2018.

Thomas, B.C., Pedersen, B. and Freeling, M. (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**, 934–946.

Throude, M., Bolot, S., Bosio, M., et al. (2009) Structure and expression analysis of rice paleo duplications. *Nucleic Acids Res.*, **37**, 1248–1259.

Tsunoyama, K., Bellgard, M.I. and Gojobori, T. (2001) Intragenic variation of synonymous substitution

rates is caused by nonrandom mutations at methylated CpG. *J. Mol. Evol.*, **53**, 456–464.

Turleau, C. and Grouchy, J. De (1973) New Observations on the Human and Chimpanzee Karyotypes. , **157**.

Ulrich, D. and Olbricht, K. (2013) Diversity of volatile patterns in sixteen Fragaria vesca L. Accessions in comparison to cultivars of Fragaria xananassa. *J. Appl. Bot. Food Qual.*, **86**, 37–46.

VanBuren, R., Man Wai, C., Wang, X., et al. (2020) Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.*, **11**, 1–11.

Vanneste, K., Baele, G., Maere, S. and Peer, Y. Van De (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous − Paleogene boundary. *Genome Res.*, **32**, 1334–1347.

Vanneste, K., Maere, S. and Peer, Y. Van de (2014) Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos. Trans. R. Soc. B Biol. Sci.*, **369**, 20130353.

Vavouri, T., Semple, J.I. and Lehner, B. (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.*, **24**, 485–488.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial Analysis of Gene Expression. *Science*, **270**, 484–487.

Venter, J.C., Smith, H.O. and Hood, L.E. (1996) A new strategy for genome sequencing. *Nature*, **381**, 364–366.

Vicient, C.M. and Casacuberta, J.M. (2017) Impact of transposable elements on polyploid plant genomes. *Ann. Bot.*, **120**, 195–207.

Vidalis, A., Živković, D., Wardenaar, R., Roquis, D., Tellier, A. and Johannes, F. (2016) Methylome evolution in plants. *Genome Biol.*, **17**, 264.

Walkowiak, S., Gao, L., Monat, C., et al. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nat. 2020*, **588**, 1–7.

Wang, B., Kumar, V., Olson, A. and Ware, D. (2019) Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. *Front. Genet.*, **10**, 1–11.

Wang, J., Sun, P., Li, Y., et al. (2017) Hierarchically aligning 10 legumegenomes establishes a family-level genomics platform. *Plant Physiol.*, **174**, 284–300.

Wang, M., Tu, L., Lin, M., et al. (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.*, **49**, 579–587.

Wang, P., Meng, F., Moore, B.M. and Shiu, S.H. (2021) Impact of short-read sequencing on the misassembly of a plant genome. *BMC Genomics*, **22**, 1–18.

Wang, X., Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L., Wendel, J.F. and Liu, B. (2017) Gene-body CG methylation and divergent expression of duplicate genes in rice. *Sci. Rep.*, **7**, 1–11.

Wang, Y., Wang, X., Lee, T.H., Mansoor, S. and Paterson, A.H. (2013) Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). *New Phytol.*, **198**, 274–283.

Wang, Y., Wang, X. and Paterson, A.H. (2012) Genome and gene duplications and gene expression divergence: A view from plants. *Ann. N. Y. Acad. Sci.*, **1256**, 1–14.

Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S.P., Feltus, F.A. and Paterson, A.H. (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across

divergent angiosperms. *PLoS One*, **6**.

Washburn, J.D. and Birchler, J.A. (2014) Polyploids as a "model system" for the study of heterosis. *Plant Reprod.*, **27**, 1–5.

Waterston, R.H., Lander, E.S. and Sulston, J.E. (2002) On the sequencing of the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 3712–3716.

Wei, N., Cronn, R., Liston, A. and Ashman, T.L. (2019) Functional trait divergence and trait plasticity confer polyploid advantage in heterogeneous environments. *New Phytol.*, **221**, 2286–2297.

Wendel, J.F. (2015) The wondrous cycles of polyploidy in plants. *Am. J. Bot.*, **102**, 1753–1756.

Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J.W., Roach, P.L., Bradley, M. and Neylon, C. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, **33**, 1–6.

Whitkus, R., Doebley, J. and Lee, M. (1992) Comparative genome mapping od Sorghum and Maize. *Genetics*, **132**, 1119–1130.

Wickland, D.P. and Hanzawa, Y. (2015) The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. *Mol. Plant*, **8**, 983–997.

Wiens, J.J. (1995) Polymorphic Characters in Phylogenetic Systematics. *Syst. Biol.*, **44**, 482–500.

Wolfe, K.H. (2001) Yesterday 's Polyploids and the mystery of diploidization. *Nat. Rev.*, **2**, 333–341.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. and Rieseberg, L.H. (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 13875–13879.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M. and Wang, X. (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci.*, **111**, 5283–5288.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. and Freeling, M. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.*, **8**.

Wu, R. and Kaiser, A.D. (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.*, **35**, 523–537.

Wu, S., Shamimuzzaman, M., Sun, H., et al. (2017) The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. *Plant J.*, **92**, 963–975.

Xu, C., Nadon, B.D., Kim, K. Do and Jackson, S.A. (2018) Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.*, **41**, 2033–2044.

Xu, J., Yuan, Y., Xu, Y., et al. (2014) Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biol.*, **14**, 1–15.

Xu, W., Zhang, Q., Yuan, W., et al. (2020) The genome evolution and low-phosphorus adaptation in white lupin. *Nat. Commun.*, **11**, 1–13.

Yang, L. and Gaut, B.S. (2011) Factors that contribute to variation in evolutionary rate among arabidopsis genes. *Mol. Biol. Evol.*, **28**, 2359–2369.

Yant, L., Hollister, J.D., Wright, K.M., Arnold, B.J., Higgins, J.D., Franklin, F.C.H. and Bomblies, K. (2013) Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr. Biol.*, **23**, 2151–2156.

Yim, W.C., Lee, B.M. and Jang, C.S. (2009) Expression diversity and evolutionary dynamics of rice duplicate genes. *Mol. Genet. Genomics*, **281**, 483–493.

Yona, A.H., Manor, Y.S., Herbst, R.H., Romano, G.H., Mitchell, A., Kupiec, M., Pilpel, Y. and Dahan, O. (2012) Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 21010–21015.

Yong, W.S., Hsu, F.M. and Chen, P.Y. (2016) Profiling genome-wide DNA methylation. *Epigenetics and Chromatin*, **9**, 1–16.

Yoo, M.-J.J., Liu, X., Pires, J.C., Soltis, P.S. and Soltis, D.E. (2014) Nonadditive gene expression in polyploids. *Annu. Rev. Genet.*, **48**, 485–517.

Yoo, M.J., Szadkowski, E. and Wendel, J.F. (2013) Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity (Edinb).*, **110**, 171–180.

Yu, X., Wang, P., Li, J., et al. (2021) Whole-Genome Sequence of Synthesized Allopolyploids in Cucumis Reveals Insights into the Genome Evolution of Allopolyploidization. *Adv. Sci.*, **8**, 1–15.

Yuan, Y., Bayer, P., Batley, J. and Edwards, D. (2021) Current status of structural variation studies in plants. *Plant Biotechnol. J.*, 1–11.

Yuan, Y., Chung, C.Y.L. and Chan, T.F. (2020) Advances in optical mapping for genomic research. *Comput. Struct. Biotechnol. J.*, **18**, 2051–2062.

Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhang, H., Lang, Z. and Zhu, J.K. (2018) Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.*, **19**, 489–506.

Zhang, J., Liu, Y., Xia, E.H., Yao, Q.Y., Liu, X.D. and Gao, L.Z. (2015) Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E7022–E7029.

Zhang, P., Li, W., Friebe, B. and Gill, B.S. (2004) Simultaneous painting of three genomes in hexaploid wheat by BAC-FISH. *Genome*, **47**, 979–987.

Zhang, Q., Guan, P., Zhao, L., et al. (2021) Asymmetric epigenome maps of subgenomes reveal imbalanced transcription and distinct evolutionary trends in Brassica napus. *Mol. Plant*, **14**, 604–619.

Zhang, X., Chen, S., Shi, L., et al. (2021) Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. *Nat. Genet.*, **53**, 1250–1259.

Zhang, X., Shiu, S., Cal, A. and Borevitz, J.O. (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genet.*, **4**.

Zhang, X., Yazaki, J., Sundaresan, A., et al. (2006) Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*, **126**, 1189–1201.

Zhao, C., Zayed, O., Zeng, F., et al. (2019) Arabinose biosynthesis is critical for salt stress tolerance in Arabidopsis. *New Phytol.*, **224**, 274–290.

Zhao, M., Zhang, B., Lisch, D. and Ma, J. (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell*, **29**, 2974–2994.

Zhao, N., Dong, Q., Nadon, B.D., Ding, X., Wang, X., Dong, Y., Liu, B., Jackson, S.A. and Xu, C. (2020) Evolution of homeologous gene expression in polyploid wheat. *Genes (Basel).*, **11**, 1–13.

Zhao, S., Ye, Z. and Stanton, R. (2020) Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, **26**, 903–909.

Zhou, Q., Tang, D., Huang, W., et al. (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.*, **52**, 1018–1023.

Zhou, X., Liao, Y., Kim, S.U., Chen, Z., Nie, G., Cheng, S., Ye, J. and Xu, F. (2020) Genome-wide identification and characterization of bHLH family genes from Ginkgo biloba. *Sci. Rep.*, **10**, 1–15.

Zhou, Y. and Gaut, B.S. (2020) Large chromosomal variants drive adaptation in sunflowers. *Nat. Plants*, **6**, 734–735.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.*, **39**, 61–69.

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B.J. and Salzberg, S.L. (2017) The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience*, **6**.

Zimin, A. V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J. and Salzberg, S.L. (2017) Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.*, **27**, 787–792.

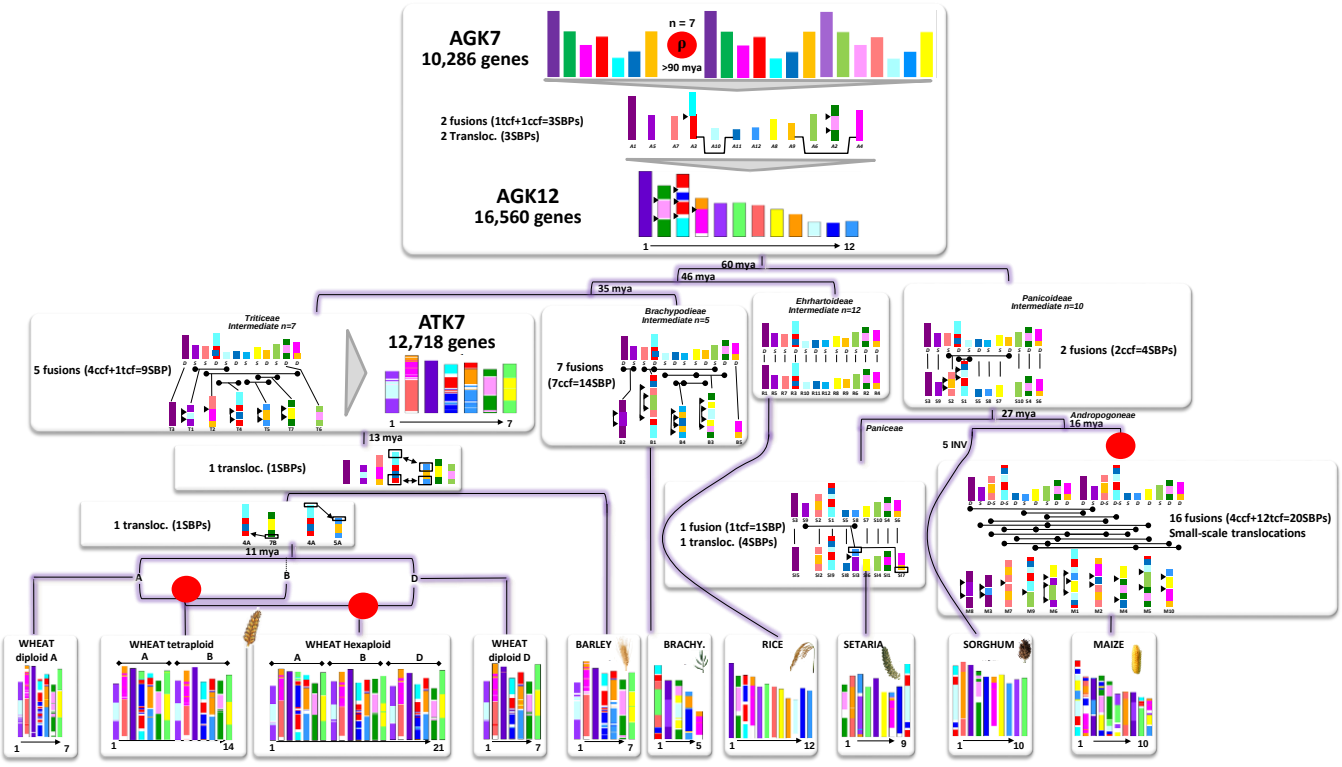
Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.*, **151**, 3–15.

V. ANNEXES

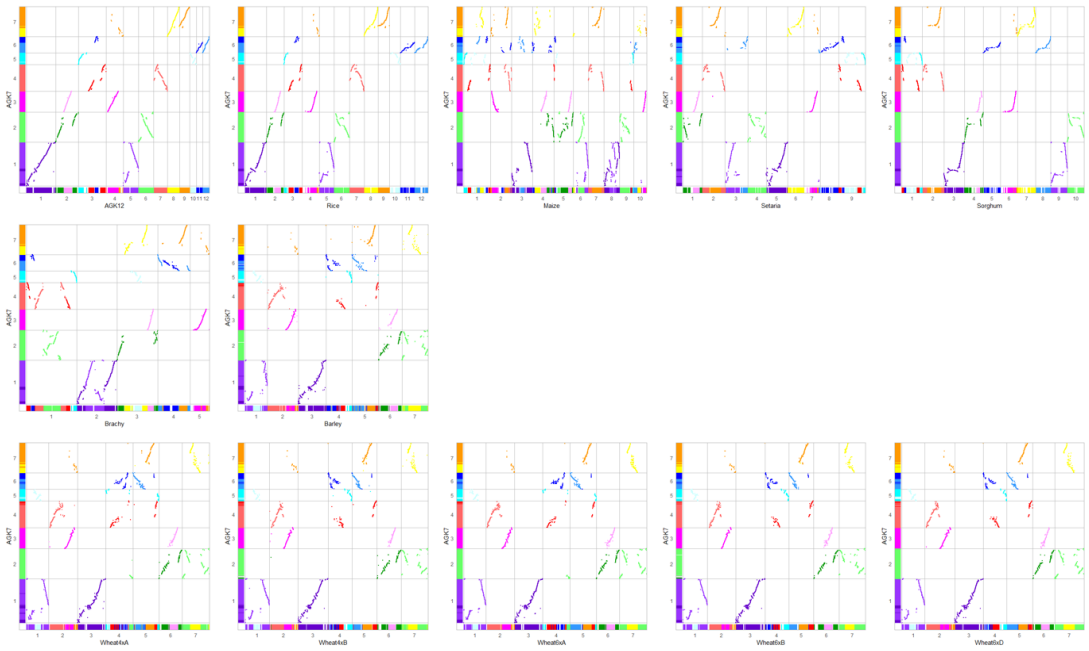
1 Figures

<u>Supplementary Figure 1</u>: Dotplot based deconvolution of modern genomes according to inferred AGKs (Ancestral Grass Karyotypes). a. Evolutionary scenario from inferred ancestors (AGK and ATK) leading to the extant grass genomes. Evolutionary events (fusions, fissions, translations) defining synteny breakpoints (SBPs) are illustrated and detailed on the tree branches. From AGK12 (structure of the modern rice genome), two ancestral chromosomes fusions (reported as ccf for centromeric-based chromosome fusions and tcf for telomeric-based chromosome fusions) contributing to 4 SBPs explain the modern sorghum genome and setaria experienced specifically an extra fusion and one complex translocation event between chromosomes 3 and 7, leading to a total of 4 and 9 SBPS in respectively setaria and sorghum. Maize experienced 16 fusions and small-scale translocations leading to 38 SBPs. *Brachypodium* experienced 14 SBPs (from 7 fusions), one common with the *Triticeae* (corresponding to a fusion between chromosomes 9 and 12 in AGK12). The modern *Triticeae* derive from a ATK7 inherited from AGK12 following 5 ancestral chromosome fusions and 2 translocations leading to 10 SBPs (and 11 for wheat subgenome A) **b**. Dotplot based deconvolution of the synteny between AGK7 (y-axis, with light and dark shades of seven colors illustrating the transition between AGK7 to AGK12) and the modern genomes (x-axis). **c**. Dotplot based deconvolution of the synteny between AGK7 (y-axis) and the modern genomes (x-axis) with blue dots for singleton genes and red dot for duplicated genes.

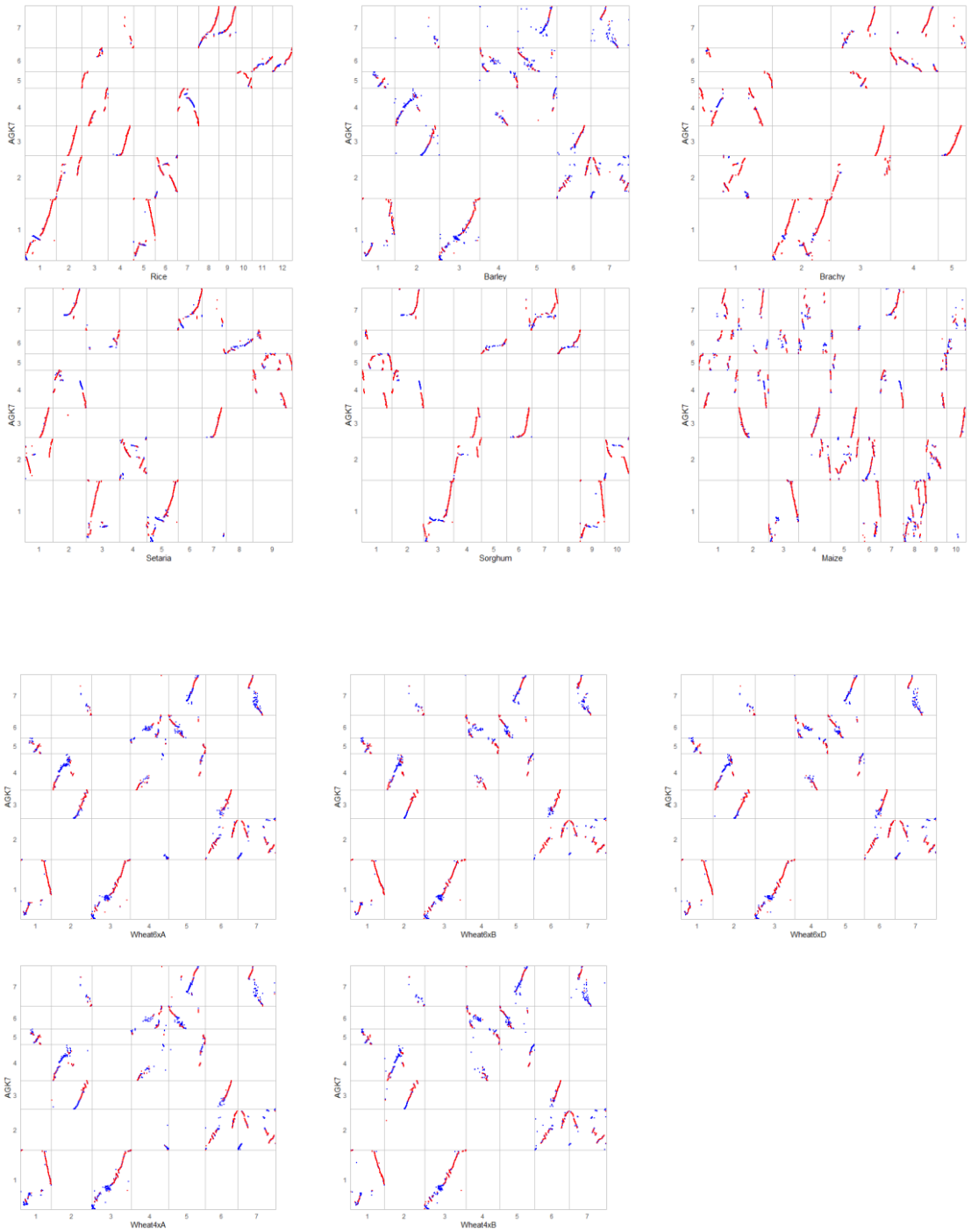
a.



b.

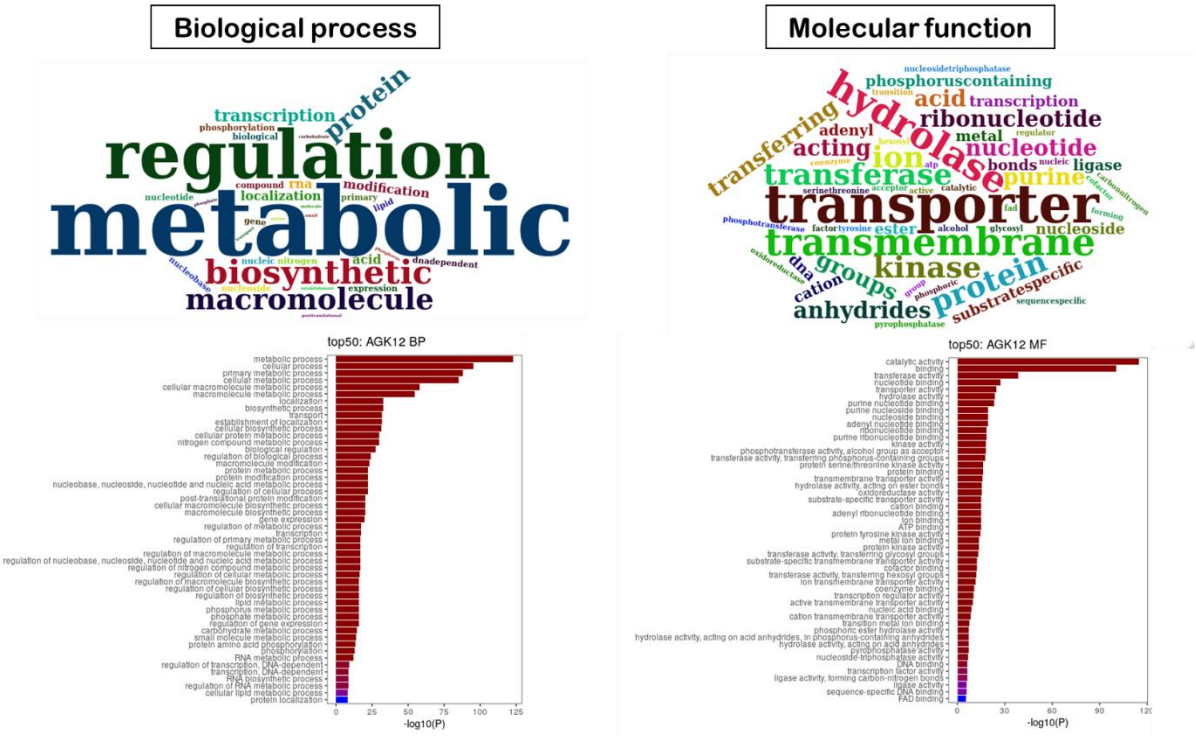

AGK7
AGK12
Rice
Maize
Setaria
Sorghum
Brachy
Barley
WheatAxA
WheatAxB
WheatBxA
WheatBxB
WheatBxD

c.

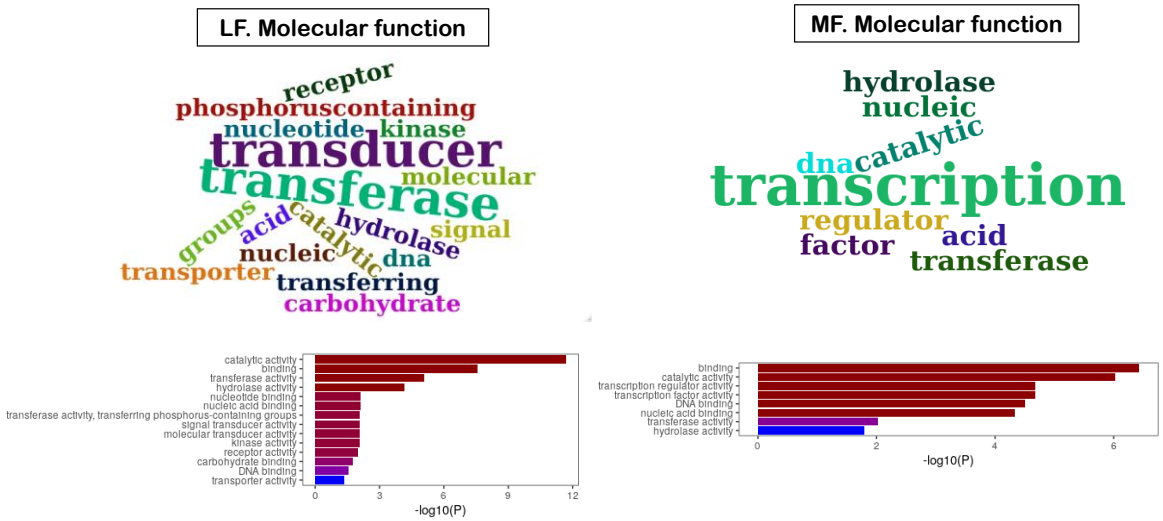

AGK7
Rice
Barley
Brachy
Setaria
Sorghum
Maize
Wheat6xA
Wheat6xB
Wheat6xD
Wheat4xA
Wheat4xB

Supplementary Figure 2: Gene Ontology enrichment. Biological processes (left) and/or molecular function (right) annotations highlighting GO enrichment presented by tag clouds and top 50 enriched GO terms for (**a**) Ancestral conserved genes in AGK12 (using rice GO as reference), (**b**) LF- and MF-genes (using rice GO as reference), (**c**) Singletons and duplicated genes pairs (using rice GO as reference), (**d**) Hypomethylated/Upregulated genes (in *Brachypodium*), (**e**) Hypermethylated/Downregulated genes (in *Brachypodium*).

a.

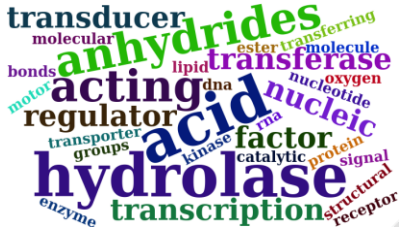


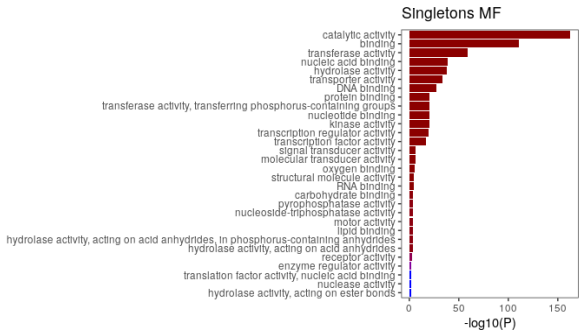
b.



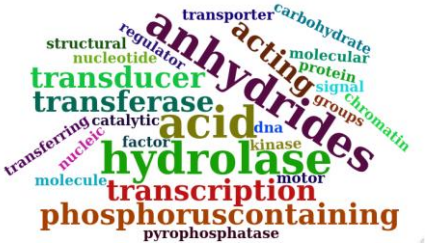
c.

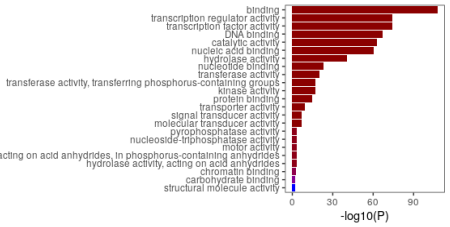
Singletons. Molecular function





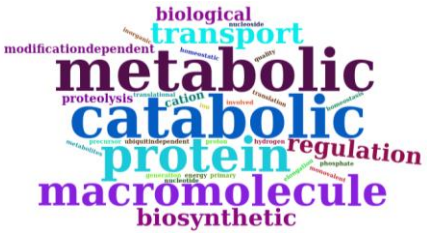
Pairs. Molecular function

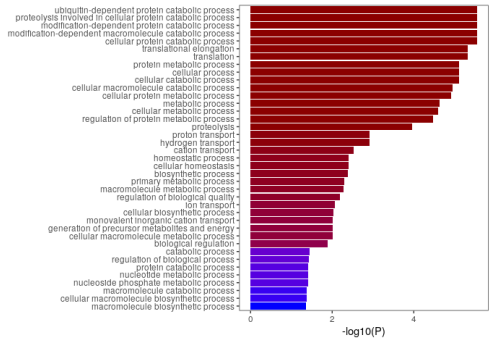




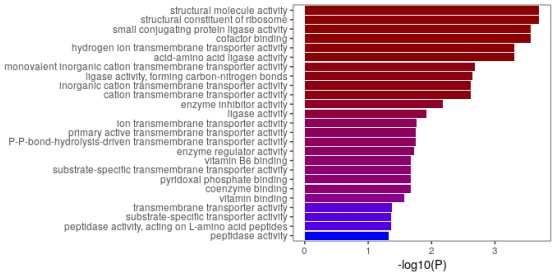
d.

Biological process



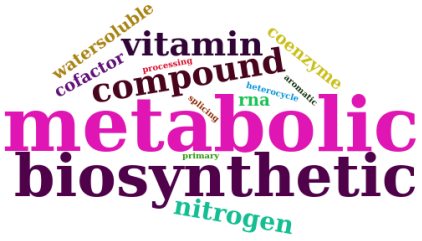


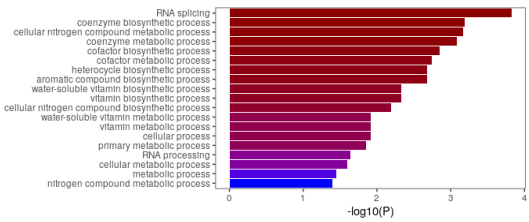
Molecular function



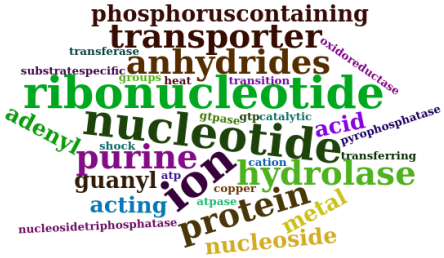
e.

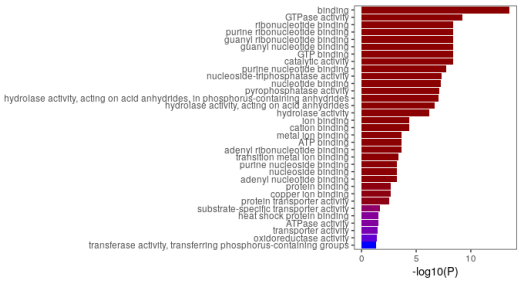
Biological process



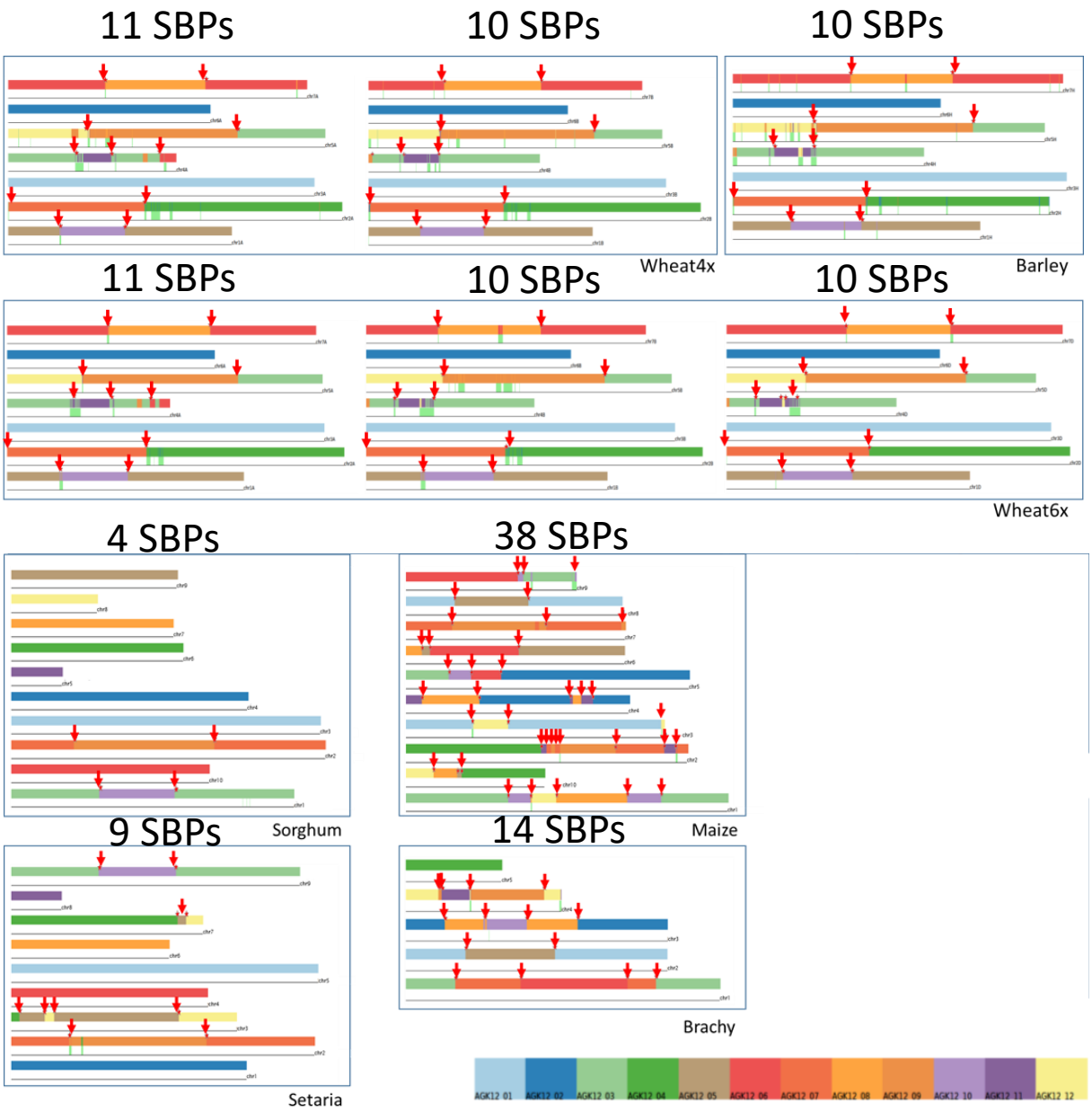


Molecular function



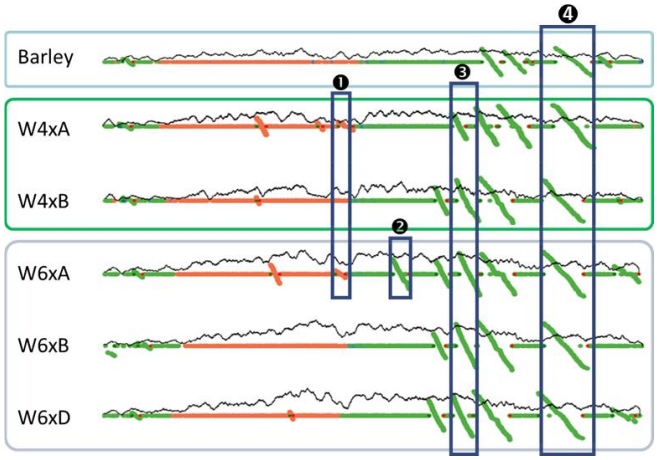


Supplementary Figure 3: Synteny break points (SBPs) between extant cereal genomes and AGK12. SBPs are illustrated by red arrows and extant chromosomes are colored according to the AGK12 color code (bottom).



Supplementary Figure 4: Inversions in grass genomes. a. Graphical representation of inversions within *Triticeae* chromosomal group 2 originating from AGK12 protochromosomes 4 (in green) and 6 (in red). Baseline represents extant gene orders when collinear to AGK12 protogenes order. Diagonals represents inversions defined as inverted extant gene orders in regards to AGK12 protogene order, with #1: inversions specific to subgenome A in wheat, #2: hexaploid wheat (6x) subgenome A specific inversions, #3: wheat specific inversions, #4. *Triticeae* specific inversions. **b**. Table illustrating the cumulative gene space covered by the inversions (in columns) detected in the investigated species (in lines) **c**. Upset plot of shared inversions among cereals. **d**. Gene density in inversions (red line) in regards to gene density resulting from 1000 simulations, by positioning the inversions randomly on the chromosomes. In all species, the observed gene density in the characterized inversions was higher than the values observed in simulations, ranging from an observed gene density higher than 87.3% of simulated values in *Brachypodium* to more than 97.9% for all other species of the panel. **e**. Position of inversions according to recombination rate levels within chromosomes.

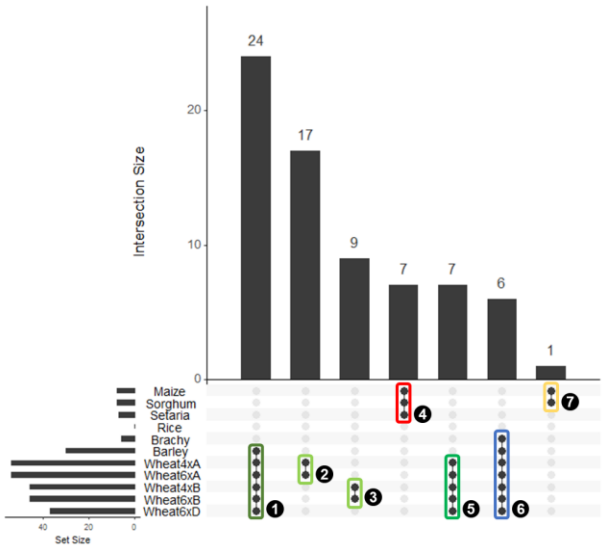
a.



b.

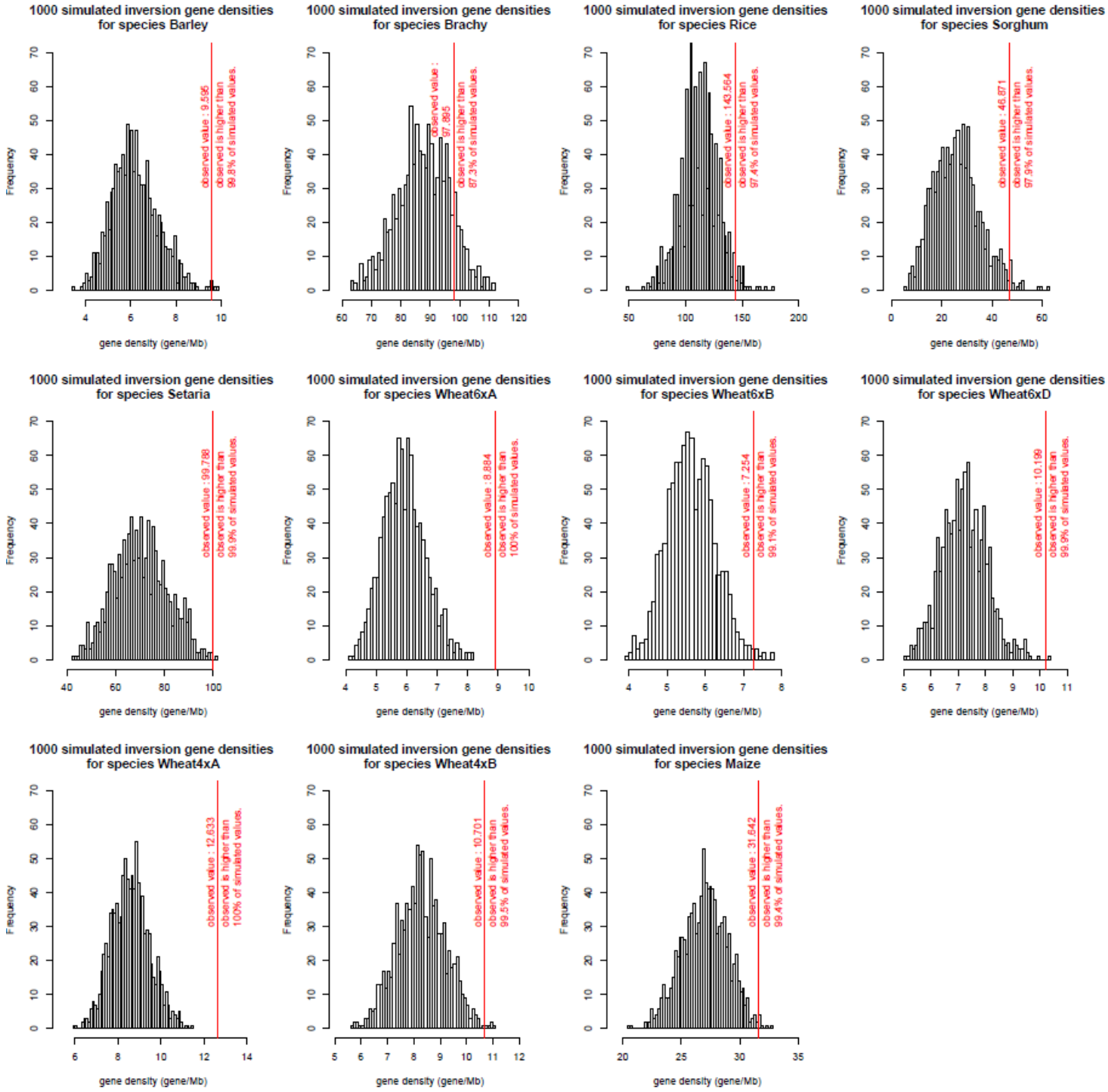
	Genome size (bp)	Cumulative inversions length (bp)	Fraction of inversions in the genome
Barley	4583636181	797115940	17,4%
Brachy	270993320	29071917	10,7%
Rice	372530218	766207	0,2%
Sorghum	659098135	82822880	12,6%
Setaria	401159754	62703147	15,6%
Wheat6xA	4934357510	1017109333	20,6%
Wheat6xB	5179939804	1145788091	22,1%
Wheat6xD	3950743238	768027870	19,4%
Wheat4xA	4899086626	1057220356	21,6%
Wheat4xB	5179230282	1182676239	22,8%
Maize	2059165538	290851024	14,1%

c.



❶: *Triticeae*-specific inversion
❷: Wheat subgenome A specific inversions
❸: Wheat subgenome B specific inversions
❹: *Panicoideae*-specific inversions
❺: Wheat specific inversions
❻: BWB specific inversions
❼: MS Specific inversions

d.

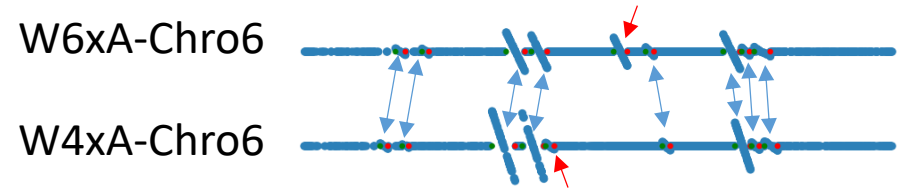

1000 simulated inversion gene densities for species Barley
observed value : 9.596
observed is higher than 99.8% of simulated values.
1000 simulated inversion gene densities for species Brachy
observed value : 97.885
observed is higher than 87.3% of simulated values.
1000 simulated inversion gene densities for species Rice
observed value : 143.564
observed is higher than 97.4% of simulated values.
1000 simulated inversion gene densities for species Sorghum
observed value : 46.871
observed is higher than 97.9% of simulated values.
1000 simulated inversion gene densities for species Setaria
observed value : 99.788
observed is higher than 99.9% of simulated values.
1000 simulated inversion gene densities for species Wheat6xA
observed value : 8.884
observed is higher than 100% of simulated values.
1000 simulated inversion gene densities for species Wheat6xB
observed value : 7.254
observed is higher than 99.1% of simulated values.
1000 simulated inversion gene densities for species Wheat6xD
observed value : 10.199
observed is higher than 99.9% of simulated values.
1000 simulated inversion gene densities for species Wheat4xA
observed value : 12.833
observed is higher than 100% of simulated values.
1000 simulated inversion gene densities for species Wheat4xB
observed value : 10.701
observed is higher than 99.5% of simulated values.
1000 simulated inversion gene densities for species Maize
observed value : 31.642
observed is higher than 99.4% of simulated values.
Frequency
gene density (gene/Mb)

e.

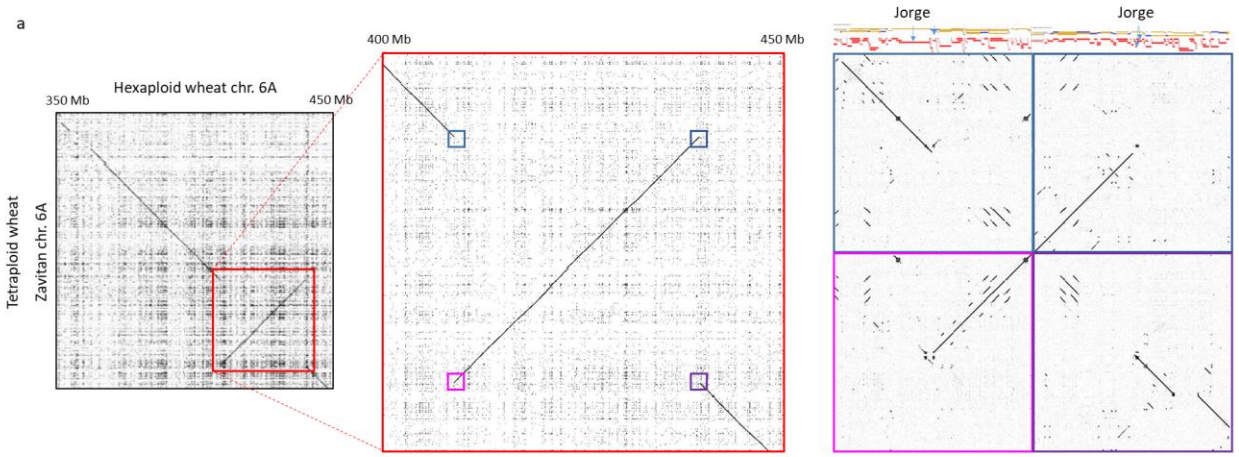
Specie	Genome Size	Cumulated LR size	Size of LR region / Total Genome Size	% of INV located in LR region
Brachy	270993320	38000000	14,0%	10,9%
Sorghum	659098135	60000000	9,1%	4,5%
Wheat6xB	5179939804	1090000000	21,0%	6,1%
Maize	2059165538	341000000	16,6%	10,5%
Barley	4583636181	1560000000	34,0%	9,8%
Setaria	401159754	70000000	17,4%	5,7%
Wheat6xD	3950743238	1040000000	26,3%	10,4%
Wheat4xA	4899086626	1780000000	36,3%	18,8%
Wheat4xB	5179230282	1180000000	22,8%	10,9%
Wheat6xA	4934357510	1720000000	34,9%	13,2%

Supplementary Figure 5: Inversions between hexaploid and tetraploid wheats. a. Detailed comparison of inversions on chromosome 6 of tetraploid and hexaploid wheat using the same representation as Supplementary Figure 4a. Blue arrows highlight shared inversions corresponding to common ancestral origins. Red arrows highlight species-specific inversions. **b**. Dotplot based alignment of Zavitan tetraploid wheat against Chinese spring hexaploid wheat of a genomic region of chromosome 6A (see coordinated on the figure). **c**. Dotplot based alignment Svevo tetraploid wheat against Chinese spring hexaploid wheat of a genomic region of chromosome 6A (see coordinated on the figure). In both cases both inversion boundaries correspond to same LTR retrotransposons (red rectangle).

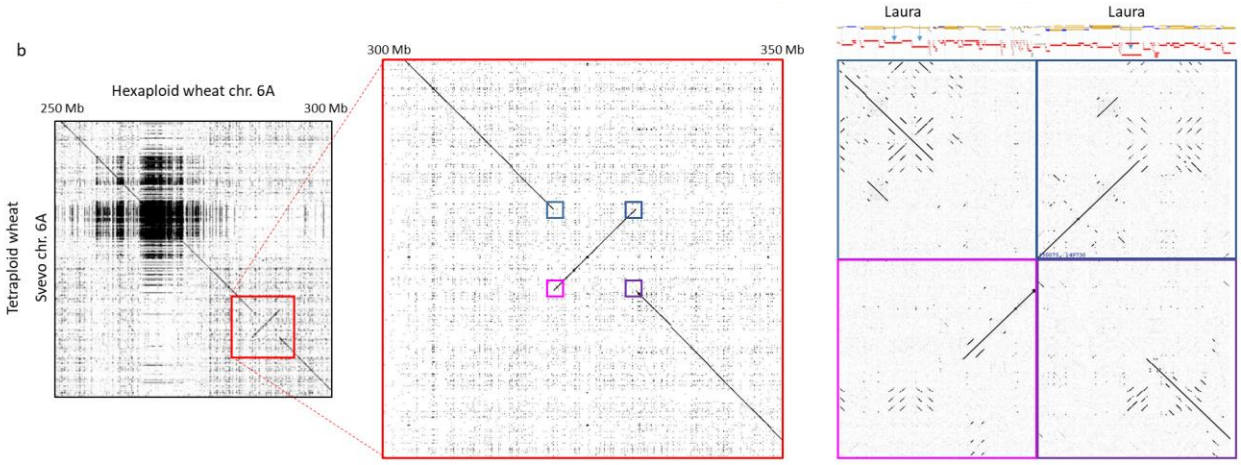
a.



b.



c.

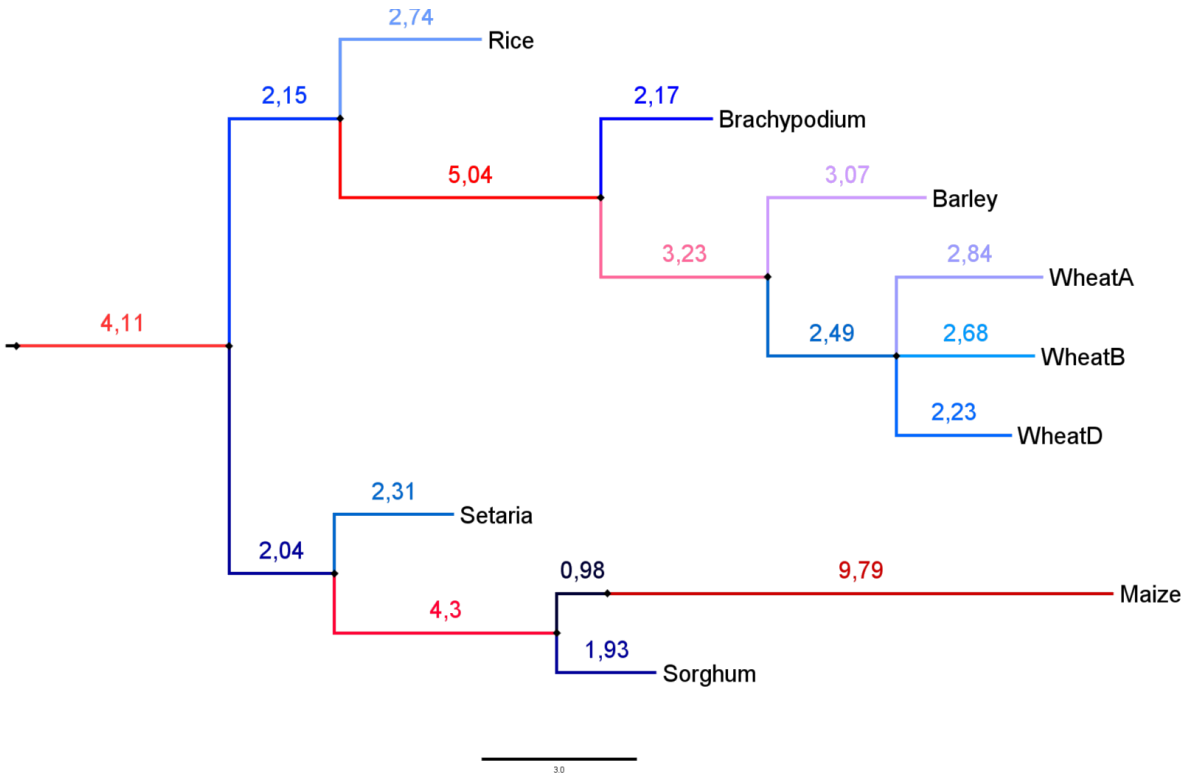


Supplementary Figure 6: Phylogeny and dating for substitution rate analysis. a. Divergence time between the different clades of cereals as estimated by Murat et al. 2017 and El Baidouri et al. 2017. **b**. Substitution rate corresponding to measured substitutions per site value normalized by branch duration in billion years among cereals. **c**. Substitution rates in *Triticeae*: the same calculation as above is applied.

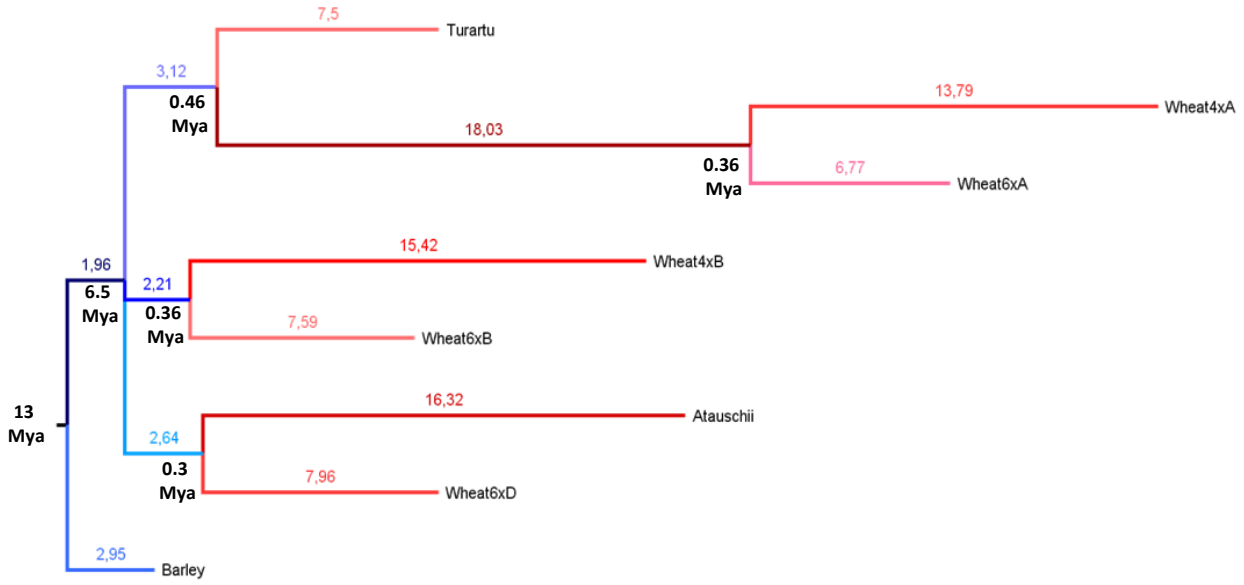
a.

Clade	Divergence time (mya)
AGK7 WGD	90
AGK12 Speciation	60
(Rice, *Brachypodium*)	46
(Barley, *Brachypodium*)	35
(Barley, Wheat)	13
Wheats root	6.5
Tetraploid wheat	0.36
Hexaploid wheat	0.01
(Maize, Setaria)	27
(Maize, Sorghum)	16
(WheatA, *T.urartu*)	0.46
(WheatD, *A.speltoides*)	0.3

b.

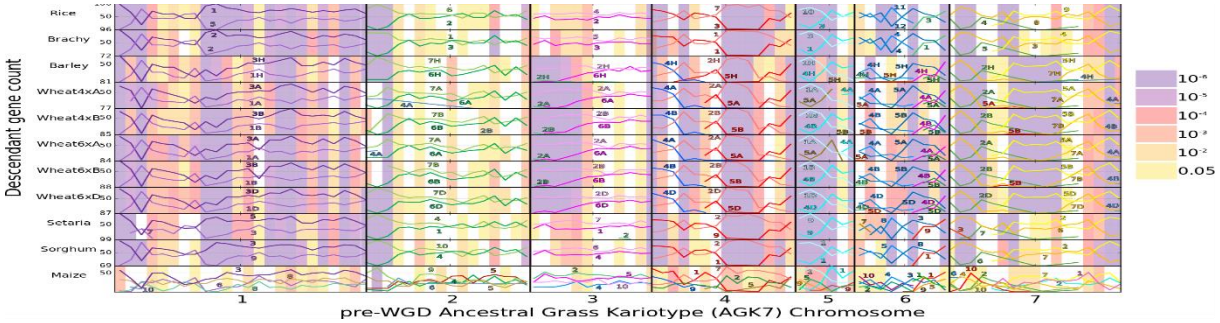


C.

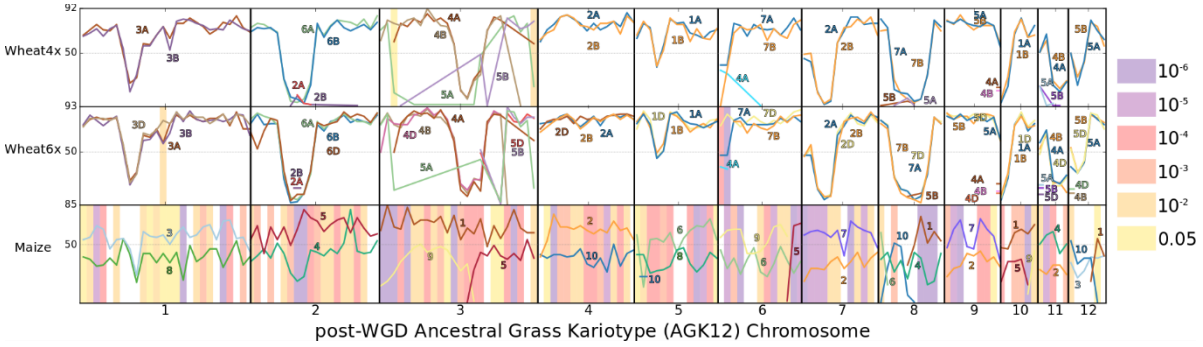

7,5
Turartu
3,12
0.46
Mya
13,79
Wheat4xA
18,03
0.36
Mya
6,77
Wheat6xA
15,42
Wheat4xB
1,96
6.5
Mya
2,21
0.36
Mya
7,59
Wheat6xB
13
Mya
16,32
Atauschii
2,64
0.3
Mya
7,96
Wheat6xD
2,95
Barley

Supplementary Figure 7: LF/MF compartments detection. a. LF/MF compartments inherited from the ancestral ρ WGD. For each ancestral chromosomes (AGK7, x-axis) curves show the counts of genes observed in the modern (post-ρ) duplicated regions in the extant species (y-axis) for each 100 ancestral genes windows. The background color indicates the p-value of biased fractionation statistical test. **b**. LF/MF compartments inherited from species-specific polyploidization events in tetraploid wheat, hexaploid wheat and maize from AGK12. Curves show the counts of genes observed in the post-duplication regions in the extant species for each 100 ancestral genes windows in the post-ρ ancestor (AGK12). The background color indicates the p-value of biased fractionation statistical test.

a.

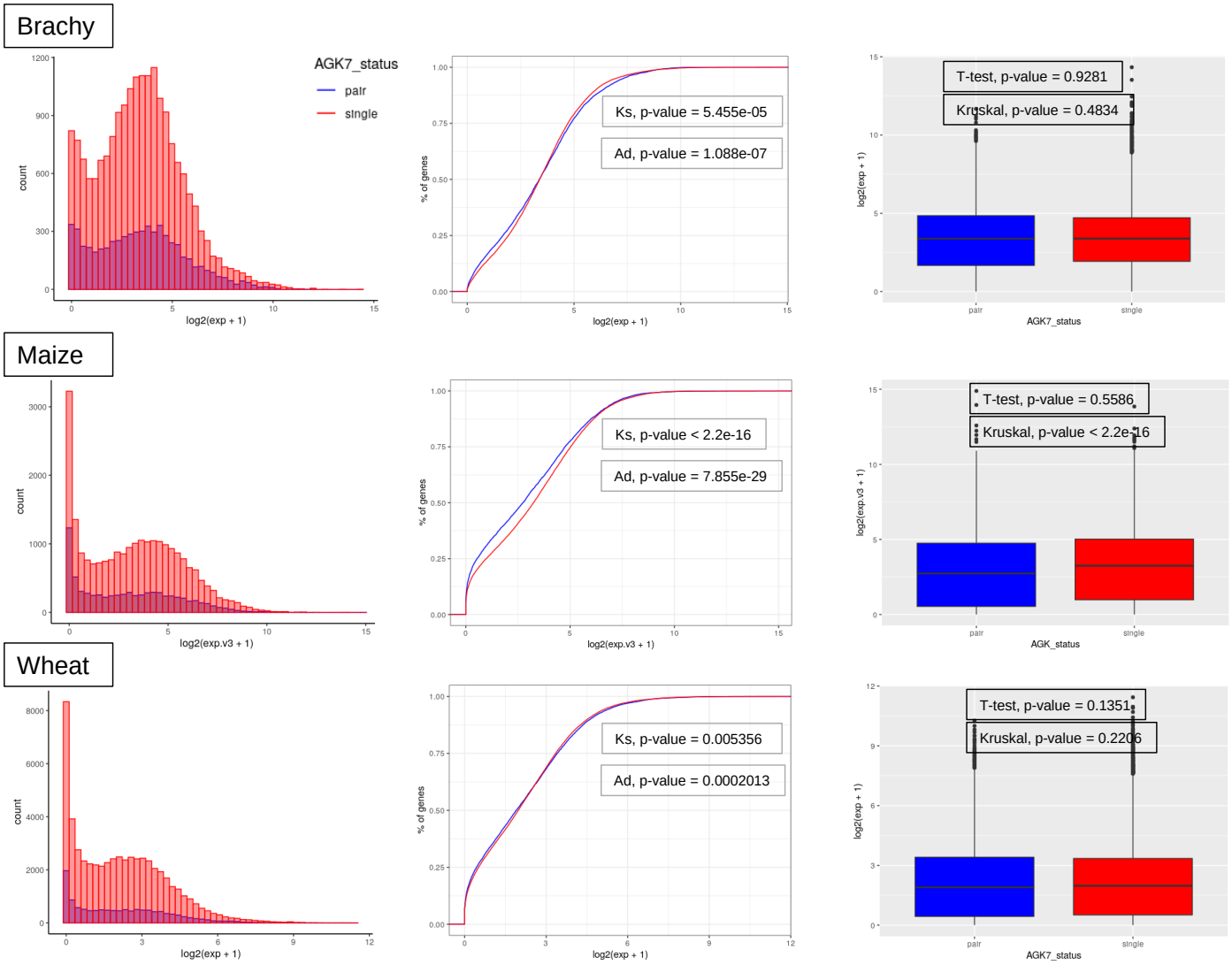


b.



Supplementary Figure 8: Analysis of omics data between collinear versus inversed genes, singletons versus genes in pairs and LF- versus MF-genes for Wheat, Maize and *Brachypodium*. a. Comparative analysis of gene expression between singletons and duplicated genes (pairs) in *Brachypodium*, maize and wheat illustrated as (left) histogram distribution plot, (center) empirical cumulative distribution and (right) boxplot. Significant differences are shown using four different statistical tests, *i.e*. Ks = Kolmogorov-Smirnov test, Ad= Anderson-Darling test, t-test= Student test, Kruskal = Kruskal-Wallis test. **b**. Data summary showing omics differences such as gene expression, DNA methylation (promoter and gene body), mutations dynamic (SNPs) and substitution rates (Ka, Ks), illustrated with arrows (↗ for increase and ↘ for decrease) between collinear and inverted, conserved and non-Conserved, singleton and pair, LF- and MF-genes. Col. = collinear; Inv. = Inverted; CS = Conserved; NCS = Non-Conserved; Sing = Singleton; pair = duplicated genes; LF = Least-Fractionated; MF = Most-Fractionated; SE = Slow evolving; FE = Fast evolving.

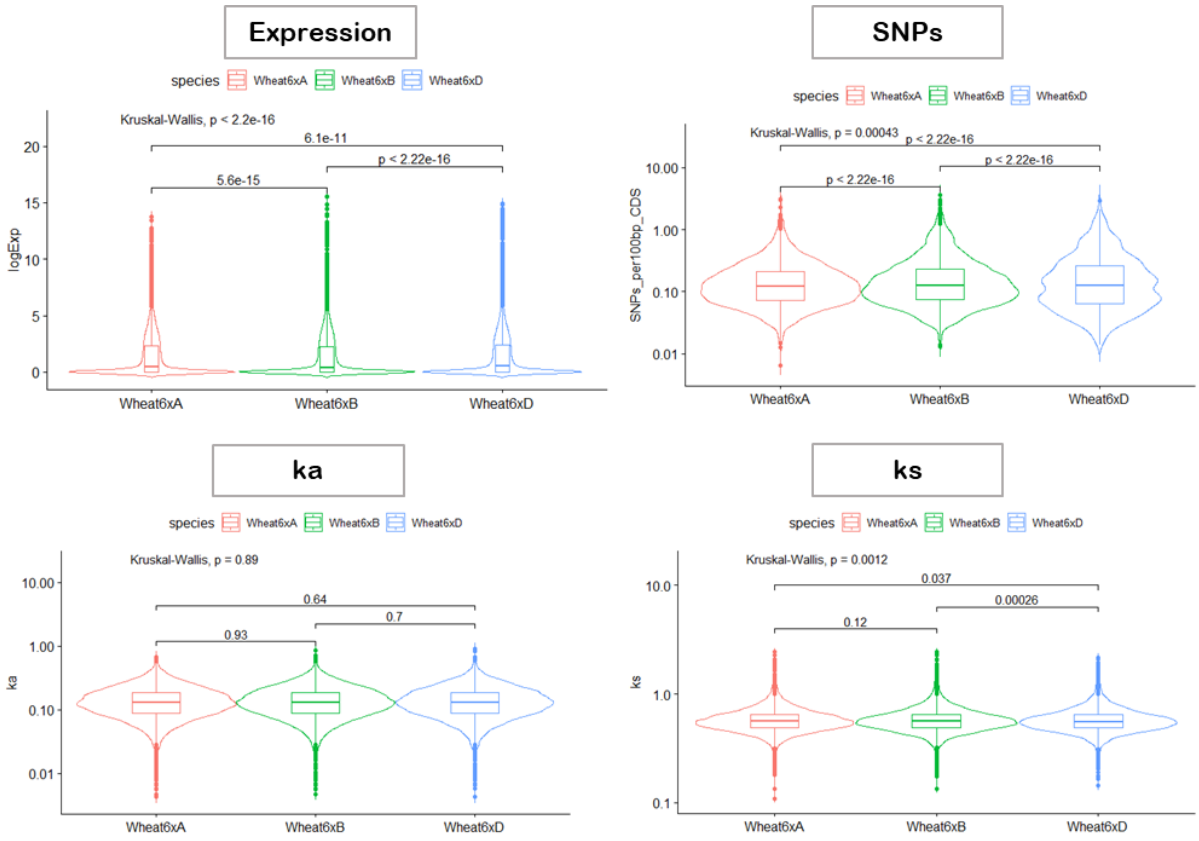
a.



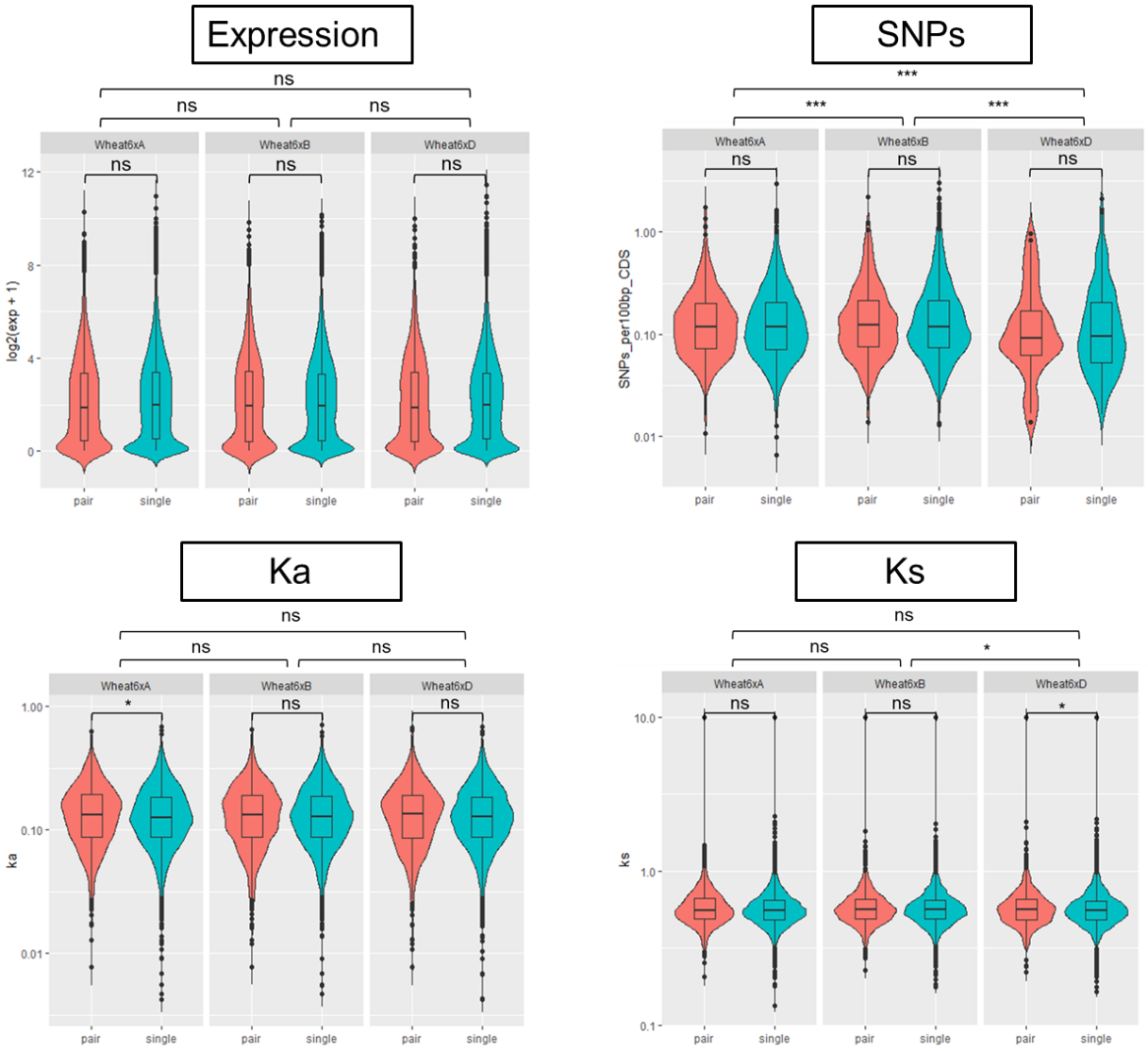
b.

		Brachy	*Brachy*	*Maize A*	*Maize A*	*Wheat6x*	*Wheat6x*	*Maize R*	*Maize R*
direction (col|inv)	Expression							X	X
	mCG promoter					X	X	X	X
	mCHG promoter					X	X	X	X
	mCHH promoter					X	X	X	X
	mCG gbM					X	X	X	X
	mCHG gbM					X	X	X	X
	mCHH gbM					X	X	X	X
	SNPs							X	X
	Ka							X	X
	Ks							X	X
Evolutionary state (**CS**|NCS)	Expression							X	X
	mCG promoter					X	X	X	X
	mCHG promoter					X	X	X	X
	mCHH promoter					X	X	X	X
	mCG gbM					X	X	X	X
	mCHG gbM					X	X	X	X
	mCHH gbM					X	X	X	X
	SNPs							X	X
	Ka							X	X
	Ks							X	X
AGK status (**sing**|**pair**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								
Fraction (**LF**|**MF**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								
Evolutionary speed (**SE**|**FE**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								

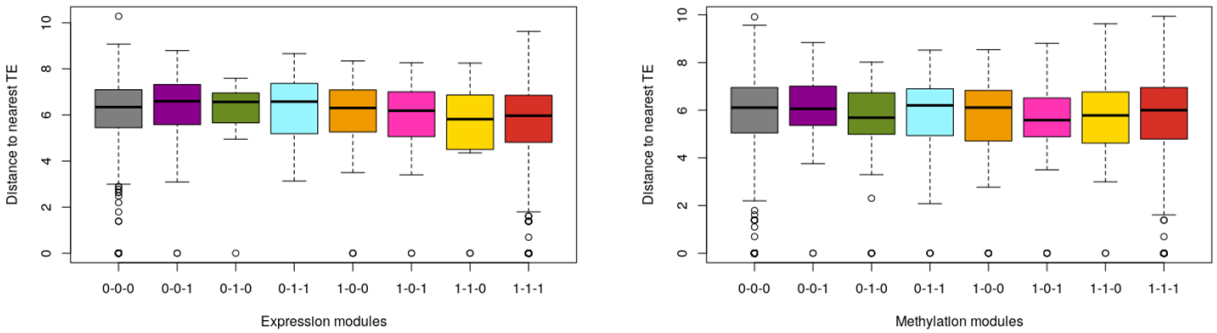
<u>Supplementary Figure 9</u>: Subgenome dominance in hexaploid wheat between A-B-D subgenomes. Comparison of gene expression (expressed as log2(exp+1)), SNP rate (expressed as SNP per 100bp in CDS), Ka and Ks values between the A, B and D subgenomes (x-axis) of hexaploid wheat. Statistical test p-value values are shown above the brackets.



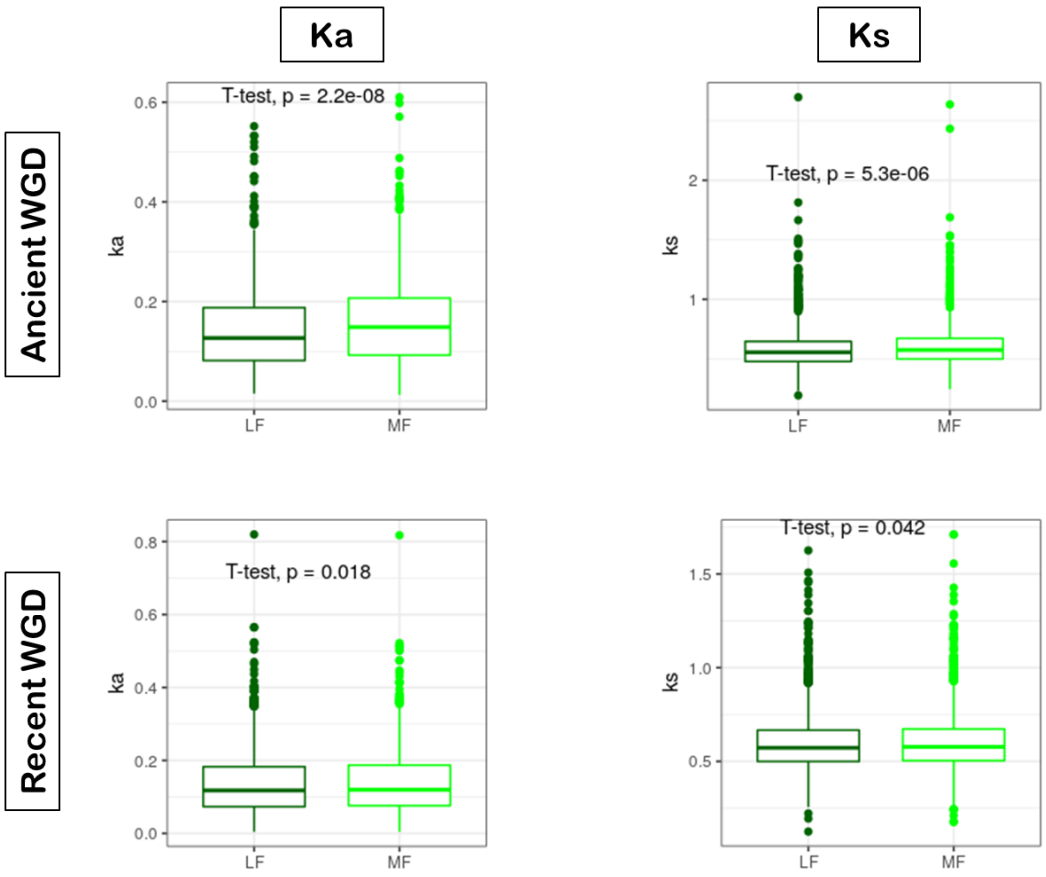
<u>Supplementary Figure 10</u>: Subgenome dominance in hexaploid wheat between recent A-B-D subgenomes for each ancient LF-MF subgenomes. Comparison of gene expression (expressed as log2(exp+1)), SNPs (expressed as SNP per 100bp in CDS), Ka and Ks values between A, B and D subgenomes of hexaploid wheat partitioned into singletons and pairs (x-axis). Values above the square brackets indicate whether the compared values show significant statistically differences. Stars indicate that the differences are significant, and their number reflects increased statistical significance. NS indicates the test is not significant. The square brackets above the boxes correspond to the comparisons between the three wheat subgenomes for a given gene category (singletons or pairs). The square brackets in the boxes correspond to comparisons between singletons and pairs within a given subgenome (either A, B or D).



<u>Supplementary Figure 11</u>: Impact of transposable elements (TEs) in methylation and expression differences observed between of conserved genes. Expression and methylation modules are represented in the x-axis and the distance to the nearest TEs (expressed in logarithm absolute value) in the y-axis. 1-1-1 corresponds to an expression observed in all the three developmental stages; 0-0-0 corresponds to genes not expressed in all the three developmental stages; Modules with both 0 and 1 represent differences in expression or methylation between the developmental stages. Zero indicates not methylated/expressed genes and 1 for methylated/expressed genes.

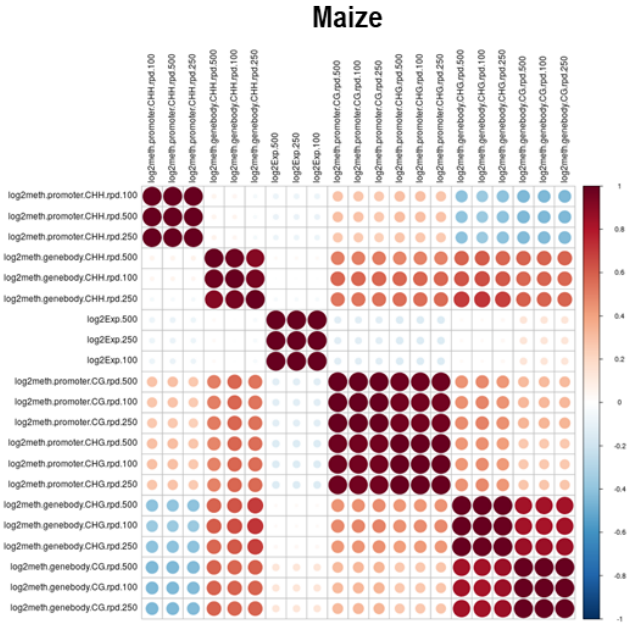


Supplementary Figure 12: Ka and Ks plasticity of duplicated genes. Non-synonymous (ka, left) and synonymous (ks, right) substitutions between the duplicated gene pairs in LF and MF compartments inherited from the ancestral shared (ρ) WGD (top) and the recent maize-specific WGD (bottom).

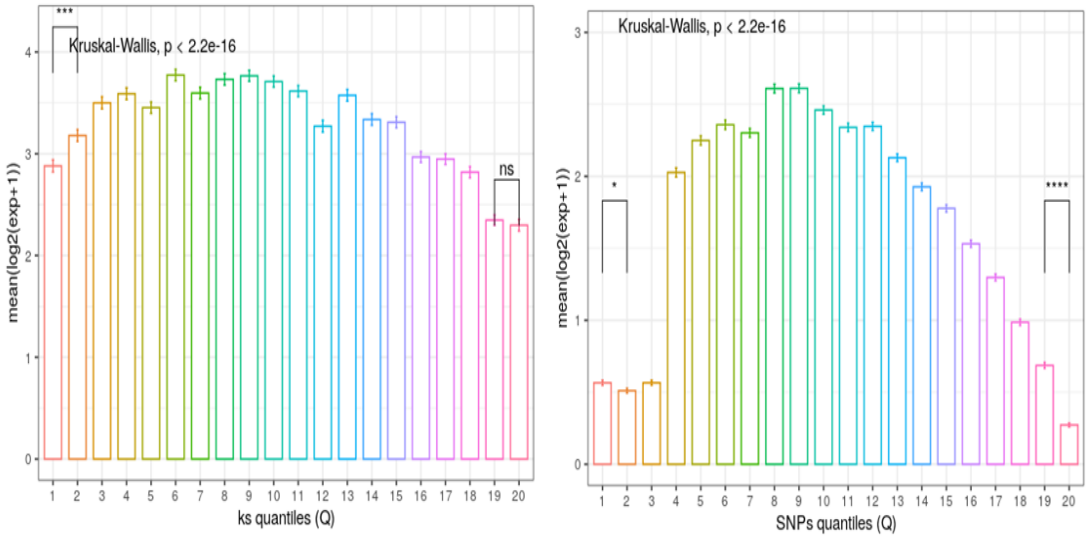


Supplementary Figure 13: Whole genome omics interplay. **a**. Whole genome correlation between genes expression and DNA methylation levels in the three contexts (CG, CHG and CHH) and for the three developmental stages (100DD, 250DD and 500DD), with red color for positive correlation and blue color for negative correlation. **b**. Interplay between gene expression level and SNP density in maize. The SNPs are represented by quantiles ranked in ascending order and gene expression by the logarithm of mean values of each quantile. **c**. Expression and methylation interplay for conserved and non-conserved genes. Methylation in gene promoter is represented in quantiles (Q1 to Q20), x-axis. Gene expression is represented as the mean value of genes in each quantile (log transformed), y-axis. Red curve for conserved genes and blue curve for non-conserved genes

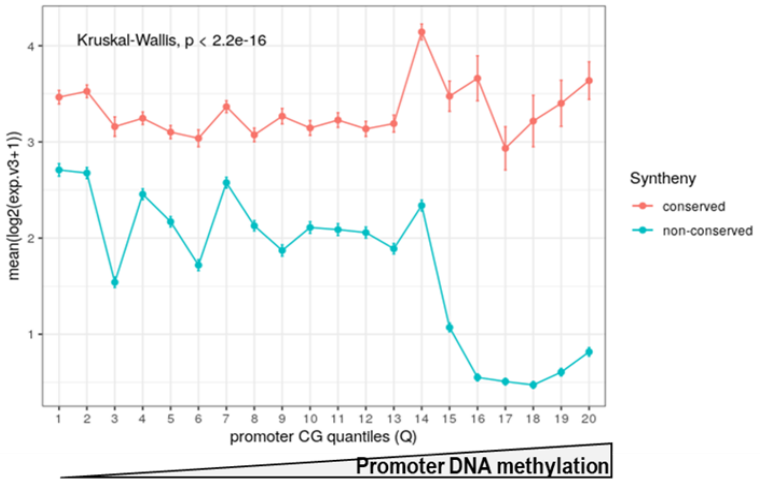
a.



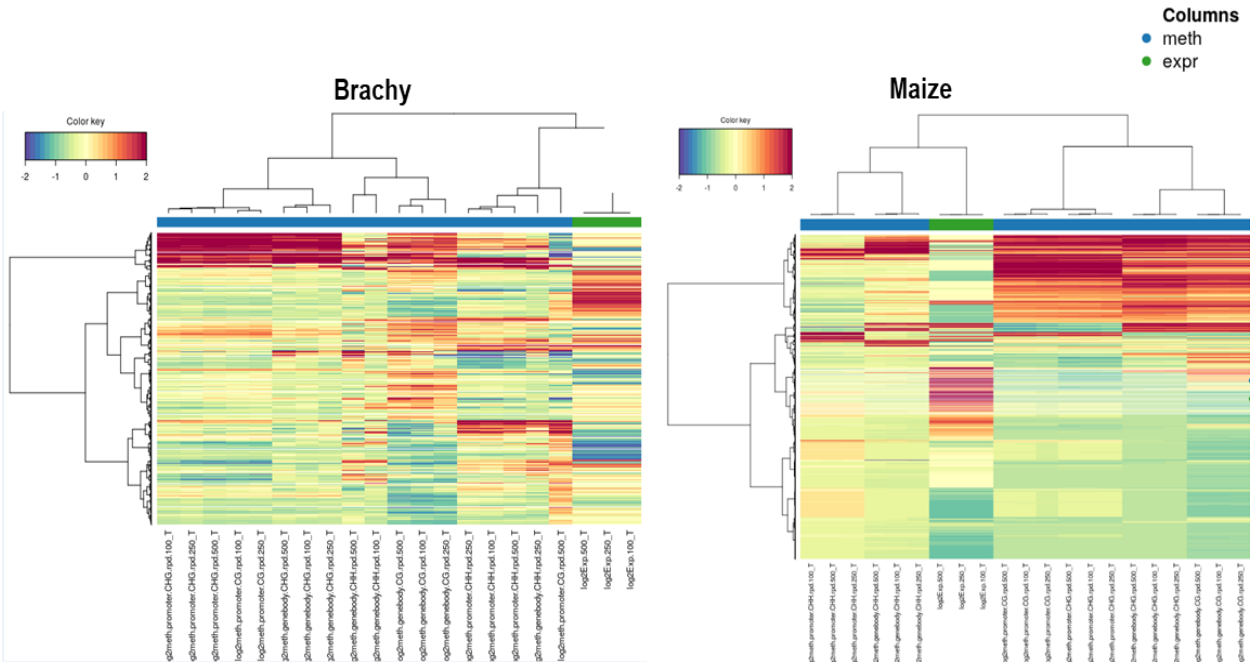
b.



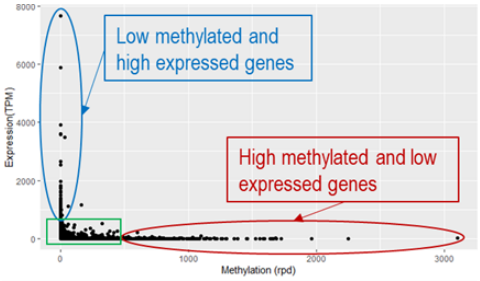
c.


Kruskal-Wallis, p < 2.2e-16
mean(log2(exp.v3+1))
promoter CG quantiles (Q)
Promoter DNA methylation
Syntheny
conserved
non-conserved

<u>Supplementary Figure 14:</u> Expression and methylation levels in Brachypodium and Maize. Mixomics data integration of (in columns) gene expression, promoter and gene-body methylation (in CG, CHG and CHH) for the three developmental stages in *Brachypodium* and maize for each of the annotated genes (in lines). Normalized expression and methylation levels of genes are shown with a color code (see legend top left)



<u>Supplementary Figure 15:</u> Expression and methylation gene outliers. ***Left panel***: Outliers detection with gene expression expressed in TPM (y-axis) and methylation expressed in rpd (read per density, x-axis). This allows the identification of hypermethyled and downregulated genes (in red) and hypomethylated and upregulated genes (in blue). ***Right panel***: Number of promoter and gene-body Hyper/Down (hypermethyled / downregulated) and Hypo/Up (hypomethylated / upregulated) genes in *Brachypodium* and maize for each methylation context (CG, CHG and CHH).



	CG		CHG		CHH	
	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up
Brachy promoter	12	146	24	130	4	132
Brachy gene body	330	11	31	11	39	11
Maize promoter	35	7	38	8	10	8
Maize gene body	252	9	265	9	4	9

2 Tables

<u>Supplementary Table 1:</u> Catalog of SBPs in cereal genomes compared to AGKs. The table lists the SBPs (lines) identified in Barley, Wheat6xA, Wheat6xB, Wheat6xD, Wheat4xA, Wheat4xB, Brachypodium, Maize, Rice, Setaria and Sorghum (first column)). Specified in the table (in columns): the species and the number of the chromosomes with SBPs, SBPs bounds positions on the extant chromosomes in base pairs (Region_start and Region_stop), the size (Region_length), the number of extant genes (Gene number), and the gene density of the SBPs (Gene density/Mb).

Species	Chromosome	Region_start	Region_end	Region_length	Gene	Gene density
Barley	1	2477071	3667228	1190157	32	26,89
Barley	1	413189358	414238236	1048878	13	12,39
Barley	2	5689544	6001783	312239	5	16,01
Barley	2	22303403	29127856	6824453	138	20,22
Barley	4	157489578	159001728	1512150	9	5,95
Barley	4	248580614	361216794	112636180	302	2,68
Barley	5	302986401	385077969	82091568	443	5,4
Barley	5	547622230	587902455	40280225	814	20,21
Barley	7	324264552	328240828	3976276	9	2,26
Barley	7	463508051	469006090	5498039	26	4,73
Brachy	1	13183945	13344688	160743	5	31,11
Brachy	1	24607778	24718606	110828	5	45,11
Brachy	1	50736361	50746905	10544	2	189,68
Brachy	1	58794709	58806320	11611	2	172,25
Brachy	2	12547382	12752760	205378	5	24,35
Brachy	2	40053404	40107177	53773	4	74,39
Brachy	3	10366965	11160532	793567	15	18,9
Brachy	3	19599028	22691094	3092066	232	75,03
Brachy	3	36917414	37033390	115976	13	112,09
Brachy	3	44389423	44398707	9284	2	215,42
Brachy	4	6543910	7166638	622728	64	102,77
Brachy	4	8167934	8388368	220434	14	63,51
Brachy	4	30433226	31962990	1529764	168	109,82
Brachy	4	43678610	43700023	21413	3	140,1
Maize	1	79336971	79740274	403303	22	54,55
Maize	1	141080871	147736435	6655564	100	15,03
Maize	1	180962581	181907817	945236	29	30,68
Maize	1	221838374	223002084	1163710	45	38,67
Maize	1	244807263	245046073	238810	9	37,69
Maize	2	113942560	121438293	7495733	145	19,34
Maize	2	147902840	157500666	9597826	223	23,23
Maize	2	161757546	164965469	3207923	76	23,69
Maize	2	167840075	167968408	128333	8	62,34
Maize	2	172156792	172462208	305416	5	16,37
Maize	2	198751071	199316619	565548	24	42,44
Maize	2	219443660	219608924	165264	15	90,76
Maize	2	231026903	233153512	2126609	119	55,96
Maize	3	67694182	86854672	19160490	255	13,31

Maize	3	143634982	143891710	256728	12	46,74
Maize	3	228749965	229646632	896667	38	42,38
Maize	4	23448577	24448923	1000346	28	27,99
Maize	4	84270643	95630081	11359438	192	16,9
Maize	4	183205800	185563721	2357921	105	44,53
Maize	4	191942375	193334572	1392197	44	31,6
Maize	4	198107413	200368095	2260682	92	40,7
Maize	4	215180538	221458660	6278122	158	25,17
Maize	5	20940827	21599690	658863	27	40,98
Maize	5	40103819	46567200	6463381	128	19,8
Maize	5	65484462	66073374	588912	16	27,17
Maize	6	9499905	25704825	16204920	350	21,6
Maize	6	46506476	54374192	7867716	121	15,38
Maize	6	123526597	123571303	44706	4	89,47
Maize	7	43581144	45558537	1977393	46	23,26
Maize	7	143860194	148212441	4352247	177	40,67
Maize	7	174908903	175221262	312359	10	32,01
Maize	8	60821906	61991768	1169862	18	15,39
Maize	8	134991206	135251399	260193	7	26,9
Maize	9	111136620	112906676	1770056	54	30,51
Maize	9	119681382	127567385	7886003	288	36,52
Maize	9	153022772	155725855	2703083	161	59,56
Maize	10	38923068	42891875	3968807	62	15,62
Maize	10	83847901	91148591	7300690	222	30,41
Setaria	2	14691861	15320349	628488	43	68,42
Setaria	2	39984933	40188449	203516	29	142,49
Setaria	3	2110559	2188907	78348	9	114,87
Setaria	3	5376387	5551743	175356	21	119,76
Setaria	3	6410768	6451263	40495	5	123,47
Setaria	3	32614231	35463469	2849238	94	32,99
Setaria	7	31579090	32520905	941815	122	129,54
Setaria	9	12967447	13324018	356571	43	120,59
Setaria	9	41081862	41237715	155853	16	102,66
Sorghum	1	16849399	17143942	294543	23	78,09
Sorghum	1	53881004	54029474	148470	12	80,82
Sorghum	2	18064093	18587787	523694	9	17,19
Sorghum	2	67436345	67443804	7459	2	268,13
Wheat4xA	1	193876562	214487820	20611258	70	3,4
Wheat4xA	1	393096216	393116123	19907	3	150,7
Wheat4xA	2	39639110	40346828	707718	23	32,5
Wheat4xA	2	427375403	429121466	1746063	10	5,73
Wheat4xA	4	224020167	410895620	186875453	616	3,3
Wheat4xA	4	534712732	535563312	850580	9	10,58
Wheat4xA	4	640594076	642609268	2015192	44	21,83
Wheat4xA	5	333245315	336426615	3181300	26	8,17
Wheat4xA	5	531500376	565010612	33510236	622	18,56
Wheat4xA	7	205288407	208290156	3001749	29	9,66
Wheat4xA	7	479128117	484857983	5729866	43	7,5
Wheat4xB	1	223634714	228760491	5125777	22	4,29
Wheat4xB	1	426538235	427332702	794467	8	10,07
Wheat4xB	2	4788282	66830059	62041777	1044	16,83
Wheat4xB	2	398426754	398882450	455696	5	10,97
Wheat4xB	4	76142892	78318398	2175506	7	3,22
Wheat4xB	4	331576523	332901244	1324721	16	12,08
Wheat4xB	5	294020242	297528667	3508425	32	9,12
Wheat4xB	5	556149212	556344677	195465	3	15,35
Wheat4xB	7	177556498	180057154	2500656	33	13,2

Wheat4xB	7	440879414	442827014	1947600	19	9,76
Wheat6xA	1	167150357	197138407	29988050	46	1,53
Wheat6xA	1	391235429	391255585	20156	2	99,23
Wheat6xA	2	34321456	36632893	2311437	41	17,74
Wheat6xA	2	431653755	433629639	1975884	5	2,53
Wheat6xA	4	227088465	363748257	136659792	211	1,54
Wheat6xA	4	542476010	552409504	9933494	90	9,06
Wheat6xA	4	641402570	688100962	46698392	495	10,6
Wheat6xA	5	286678678	338896927	52218249	305	5,84
Wheat6xA	5	569793013	569806330	13317	2	150,18
Wheat6xA	7	206588244	211797405	5209161	47	9,02
Wheat6xA	7	482325549	490165650	7840101	50	6,38
Wheat6xB	1	217283351	249238793	31955442	48	1,5
Wheat6xB	1	420994227	421095535	101308	2	19,74
Wheat6xB	2	49487867	55885914	6398047	80	12,5
Wheat6xB	2	396925443	397011549	86106	3	34,84
Wheat6xB	4	68095055	80991637	12896582	96	7,44
Wheat6xB	4	230008346	357960778	127952432	161	1,26
Wheat6xB	5	288175906	289164785	988879	8	8,09
Wheat6xB	5	550488462	550562162	73700	2	27,14
Wheat6xB	7	161514487	164103745	2589258	18	6,95
Wheat6xB	7	433297024	435003806	1706782	14	8,2
Wheat6xD	1	146490773	156136409	9645636	14	1,45
Wheat6xD	1	310814772	311361747	546975	4	7,31
Wheat6xD	2	32798475	33723620	925145	17	18,38
Wheat6xD	2	327067861	327758675	690814	4	5,79
Wheat6xD	4	53862639	54682312	819673	9	10,98
Wheat6xD	4	155574671	279298186	123723515	167	1,35
Wheat6xD	5	254586311	255579380	993069	10	10,07
Wheat6xD	5	450577454	450636728	59274	2	33,74
Wheat6xD	7	196906407	199563813	2657406	24	9,03
Wheat6xD	7	417612156	425117899	7505743	69	9,19

Supplementary Table 2: Catalog of inversions (INVs) in cereal genomes compared to AGKs. The table lists the inversions (lines) identified in Barley, Wheat6xA, Wheat6xB, Wheat6xD, Wheat4xA, Wheat4xB, Brachypodium, Maize, Rice, Setaria and Sorghum (first column). Specified in the table (in columns): the species and the number of the chromosomes with inversions, inversions bounds positions on the extant chromosomes in base pairs (Region_start and Region_stop), the size (Region_length), the number of extant genes (Gene number) and the gene density of the inversions (Gene density/Mb).

Species	Chromosome	Region_start	Region_end	Region_lengt	Extant_nbGene	Gene
Barley	1	316925148	322794849	5869701	38	6,47
Barley	1	346195730	352077714	5881984	26	4,42
Barley	1	402304067	418582709	16278642	129	7,92
Barley	1	418909918	437346341	18436423	166	9
Barley	1	440111842	454083778	13971936	122	8,73
Barley	1	456416415	463803517	7387102	69	9,34
Barley	1	475310599	507980869	32670270	414	12,67
Barley	1	514065179	519000375	4935196	93	18,84
Barley	1	524009316	527008455	2999139	71	23,67
Barley	2	11225540	13270008	2044468	68	33,26

Barley	2	659264340	679394981	20130641	213	10,58
Barley	2	682412398	698444405	16032007	271	16,9
Barley	2	698876861	703955368	5078507	95	18,71
Barley	2	715709894	738561125	22851231	415	18,16
Barley	2	743376231	753304022	9927791	218	21,96
Barley	2	754872592	757677154	2804562	49	17,47
Barley	3	39982499	48159474	8176975	97	11,86
Barley	3	49667067	95373414	45706347	297	6,5
Barley	3	102203743	117877676	15673933	80	5,1
Barley	3	118884357	134493168	15608811	81	5,19
Barley	3	595204527	623228705	28024178	391	13,95
Barley	3	625413741	634078800	8665059	145	16,73
Barley	4	12470902	33451633	20980731	296	14,11
Barley	4	545803574	551359371	5555797	31	5,58
Barley	4	586245141	608380775	22135634	342	15,45
Barley	4	609237274	614657915	5420641	86	15,87
Barley	4	625193355	628373278	3179923	63	19,81
Barley	5	480242588	506796785	26554197	294	11,07
Barley	5	587971902	592329607	4357705	78	17,9
Barley	6	112006630	139754711	27748081	171	6,16
Barley	6	142400821	174773122	32372301	161	4,97
Barley	6	181328904	195457851	14128947	83	5,87
Barley	6	514097241	527832437	13735196	150	10,92
Barley	6	529084151	533617612	4533461	60	13,23
Barley	6	536280417	542717101	6436684	99	15,38
Barley	7	153297246	161417025	8119779	54	6,65
Barley	7	279655007	463242433	183587426	679	3,7
Barley	7	514973657	543665852	28692195	222	7,74
Barley	7	564964839	577793427	12828588	125	9,74
Barley	7	578964313	641655637	62691324	960	15,31
Barley	7	649999627	654902054	4902427	134	27,33
Brachy	1	7866789	8571786	704997	72	102,13
Brachy	1	14733706	14914154	180448	30	166,25
Brachy	1	16729430	17213694	484264	62	128,03
Brachy	1	25406806	25503236	96430	18	186,66
Brachy	1	25530411	25577948	47537	11	231,4
Brachy	1	25725233	30732744	5007511	506	101,05
Brachy	1	34630693	35679219	1048526	51	48,64
Brachy	1	38392186	39762906	1370720	60	43,77
Brachy	1	48170520	48745122	574602	77	134,01
Brachy	1	48844141	49312683	468542	72	153,67
Brachy	1	49489327	49670019	180692	20	110,69
Brachy	1	56323254	57327959	1004705	101	100,53
Brachy	1	66748063	66951241	203178	25	123,04
Brachy	1	67402588	67768506	365918	37	101,12
Brachy	2	13336971	13672936	335965	27	80,37
Brachy	2	15276388	15434129	157741	28	177,51
Brachy	2	25460934	25714091	253157	23	90,85
Brachy	2	30057472	31050030	992558	50	50,37
Brachy	2	37475805	38930794	1454989	164	112,72
Brachy	3	2219470	2327820	108350	17	156,9
Brachy	3	9153139	9836906	683767	51	74,59
Brachy	3	13516274	15252218	1735944	203	116,94
Brachy	3	19555617	20733057	1177440	107	90,88
Brachy	3	20803942	21547714	743772	43	57,81
Brachy	3	22737292	23610785	873493	21	24,04
Brachy	3	32597331	32995339	398008	34	85,43

Brachy	3	37115681	37518794	403113	48	119,07
Brachy	3	37556700	37689382	132682	18	135,66
Brachy	3	39171286	39445203	273917	36	131,43
Brachy	3	40498283	40796240	297957	32	107,4
Brachy	3	42502915	42780295	277380	36	129,79
Brachy	3	43945969	44039666	93697	18	192,11
Brachy	3	52204981	52442048	237067	31	130,76
Brachy	3	56151878	56470424	318546	57	178,94
Brachy	3	56515587	56730506	214919	35	162,85
Brachy	3	56748178	59072840	2324662	292	125,61
Brachy	4	3253049	3520740	267691	30	112,07
Brachy	4	3687248	3826824	139576	22	157,62
Brachy	4	22156198	22516681	360483	22	61,03
Brachy	5	794061	2604362	1810301	150	82,86
Brachy	5	11300166	11627570	327404	19	58,03
Brachy	5	14589856	14937873	348017	40	114,94
Brachy	5	23482100	23619744	137644	18	130,77
Brachy	5	25883388	25984858	101470	16	157,68
Brachy	5	26139974	26307818	167844	25	148,95
Brachy	5	27507967	27692260	184293	30	162,78
Maize	1	20901476	24931263	4029787	23	5,71
Maize	1	29152720	30895201	1742481	34	19,51
Maize	1	52343963	52918870	574907	23	40,01
Maize	1	98344349	100357267	2012918	36	17,88
Maize	1	189719492	191483144	1763652	41	23,25
Maize	1	197496223	206310775	8814552	69	7,83
Maize	1	216633236	218853569	2220333	64	28,82
Maize	1	239567104	240649230	1082126	19	17,56
Maize	1	250129175	255385569	5256394	21	4
Maize	1	258156480	260427079	2270599	149	65,62
Maize	2	21631613	22129615	498002	27	54,22
Maize	2	24987243	26763090	1775847	69	38,85
Maize	2	164961363	167081632	2120269	50	23,58
Maize	2	185666362	186503354	836992	103	123,06
Maize	2	236266508	237642179	1375671	159	115,58
Maize	3	1260778	3127454	1866676	51	27,32
Maize	3	3342194	4335770	993576	55	55,36
Maize	3	5429600	8757195	3327595	77	23,14
Maize	3	8891773	19956337	11064564	23	2,08
Maize	3	20067447	28618723	8551276	117	13,68
Maize	3	38600230	40936459	2336229	39	16,69
Maize	3	86916918	94153726	7236808	32	4,42
Maize	3	113364679	123805591	10440912	177	16,95
Maize	3	137813758	138910797	1097039	25	22,79
Maize	3	196936142	198678240	1742098	30	17,22
Maize	3	218867408	225636373	6768965	72	10,64
Maize	4	10038766	11904065	1865299	51	27,34
Maize	4	21158461	23377046	2218585	67	30,2
Maize	4	41064110	42174905	1110795	40	36,01
Maize	4	47969606	54130321	6160715	51	8,28
Maize	4	64801561	71937001	7135440	110	15,42
Maize	4	185559979	190082746	4522767	68	15,04
Maize	4	204471126	211373398	6902272	70	10,14
Maize	4	212694457	215185303	2490846	61	24,49
Maize	4	238872881	239988267	1115386	27	24,21
Maize	5	41388575	43180945	1792370	40	22,32
Maize	5	105681405	124002436	18321031	198	10,81

Maize	5	129258722	135236032	5977310	114	19,07
Maize	5	135878681	144322782	8444101	68	8,05
Maize	5	145254488	156484687	11230199	64	5,7
Maize	6	27897698	41352250	13454552	63	4,68
Maize	6	116206196	118279026	2072830	57	27,5
Maize	6	123615202	126869598	3254396	103	31,65
Maize	6	127773579	129657893	1884314	107	56,78
Maize	6	129975717	130430641	454924	67	147,28
Maize	6	131324255	132848899	1524644	202	132,49
Maize	6	137492993	140497751	3004758	145	48,26
Maize	6	159455359	160898945	1443586	287	198,81
Maize	6	163118403	163425833	307430	78	253,72
Maize	6	167207086	168302184	1095098	73	66,66
Maize	7	144165640	144796986	631346	197	312,03
Maize	7	145262462	147419875	2157413	121	56,09
Maize	7	150044729	150699452	654723	207	316,16
Maize	8	17505	6360557	6343052	171	26,96
Maize	8	138202114	139129894	927780	201	216,65
Maize	8	155999788	156638463	638675	310	485,38
Maize	8	156777995	158110215	1332220	292	219,18
Maize	8	158560400	162337076	3776676	97	25,68
Maize	8	162341426	162835037	493611	336	680,7
Maize	8	163067685	166362143	3294458	63	19,12
Maize	8	166395550	167725004	1329454	194	145,92
Maize	8	168473111	171977242	3504131	248	70,77
Maize	8	172795693	175252414	2456721	185	75,3
Maize	9	61459900	70427758	8967858	271	30,22
Maize	9	88137603	89741950	1604347	159	99,11
Maize	9	93426568	103504884	10078316	137	13,59
Maize	9	106428530	108544101	2115571	150	70,9
Maize	9	121136691	125218290	4081599	129	31,61
Maize	9	144785098	145552724	767626	145	188,89
Maize	9	151990158	153087623	1097465	255	232,35
Maize	10	1948805	4568012	2619207	235	89,72
Maize	10	16577080	18869396	2292316	84	36,64
Maize	10	22133369	29634878	7501509	272	36,26
Maize	10	55997722	71066861	15069139	267	17,72
Maize	10	92793936	95782866	2988930	441	147,54
Maize	10	134741566	139284531	4542965	263	57,89
Rice	3	33790894	33907614	116720	16	137,08
Rice	4	24560986	24660068	99082	12	121,11
Rice	5	940426	1061998	121572	22	180,96
Rice	5	27463147	27636451	173304	25	144,26
Rice	8	23730714	23876240	145526	22	151,18
Rice	9	20641465	20751468	110003	17	154,54
Setaria	1	608529	9891069	9282540	1086	116,99
Setaria	2	20177286	22239544	2062258	71	34,43
Setaria	2	23425435	24188385	762950	38	49,81
Setaria	2	34102427	34466173	363746	29	79,73
Setaria	2	40888234	41112305	224071	34	151,74
Setaria	2	47797991	47945554	147563	21	142,31
Setaria	3	41149	1234577	1193428	204	170,94
Setaria	3	2504990	3813306	1308316	200	152,87
Setaria	3	4914219	5369879	455660	67	147,04
Setaria	3	5549029	6397152	848123	120	141,49
Setaria	3	6447471	7351538	904067	108	119,46
Setaria	3	7393847	7679945	286098	34	118,84

Setaria	3	11062136	11262910	200774	31	154,4
Setaria	3	40881465	46519857	5638392	385	68,28
Setaria	3	48841546	49674046	832500	99	118,92
Setaria	4	2589891	5761977	3172086	358	112,86
Setaria	4	5952609	7111619	1159010	106	91,46
Setaria	4	30287018	37416384	7129366	722	101,27
Setaria	4	37730579	38585669	855090	114	133,32
Setaria	5	111335	2157486	2046151	203	99,21
Setaria	5	2179534	5538626	3359092	440	130,99
Setaria	5	25007754	25529909	522155	42	80,44
Setaria	5	31614513	32223278	608765	56	91,99
Setaria	6	1267500	8039916	6772416	689	101,74
Setaria	6	21778836	23556354	1777518	57	32,07
Setaria	6	32835061	33142014	306953	47	153,12
Setaria	6	35193795	35509578	315783	51	161,5
Setaria	7	23571658	23719331	147673	19	128,66
Setaria	7	31579090	33157093	1578003	236	149,56
Setaria	7	33349501	33948532	599031	96	160,26
Setaria	8	8008537	13312829	5304292	278	52,41
Setaria	9	2372680	2455940	83260	18	216,19
Setaria	9	20302518	21858768	1556250	117	75,18
Setaria	9	50158230	50680372	522142	79	151,3
Setaria	9	51986679	52364304	377625	35	92,68
Sorghum	1	1961892	2117299	155407	19	122,26
Sorghum	1	3186555	3362611	176056	26	147,68
Sorghum	1	17142672	47862843	30720171	700	22,79
Sorghum	1	47990059	50294259	2304200	129	55,98
Sorghum	1	50559393	53446741	2887348	180	62,34
Sorghum	1	53587929	53876356	288427	31	107,48
Sorghum	1	65054867	65539890	485023	43	88,66
Sorghum	2	661375	809775	148400	19	128,03
Sorghum	2	68289845	68518346	228501	27	118,16
Sorghum	2	77217246	77361137	143891	17	118,14
Sorghum	3	25612	9559235	9533623	901	94,51
Sorghum	3	49630796	50717368	1086572	32	29,45
Sorghum	4	57145498	63716820	6571322	678	103,18
Sorghum	5	10631712	19218797	8587085	187	21,78
Sorghum	6	3829174	5964350	2135176	50	23,42
Sorghum	6	32028615	36787553	4758938	31	6,51
Sorghum	6	50385389	50540402	155013	17	109,67
Sorghum	7	12892714	14511779	1619065	16	9,88
Sorghum	7	58360754	64305272	5944518	604	101,61
Sorghum	8	1254991	2937703	1682712	155	92,11
Sorghum	9	56793633	56954015	160382	21	130,94
Sorghum	10	20515882	23566932	3051050	16	5,24
Wheat4xA	1	89986734	112197612	22210878	153	6,89
Wheat4xA	1	117470501	131104185	13633684	158	11,59
Wheat4xA	1	295294731	302203456	6908725	59	8,54
Wheat4xA	1	302767060	307831419	5064359	38	7,5
Wheat4xA	1	320756857	326709239	5952382	49	8,23
Wheat4xA	1	335561575	344123587	8562012	78	9,11
Wheat4xA	1	378058101	395205210	17147109	164	9,56
Wheat4xA	1	402889249	406454122	3564873	30	8,42
Wheat4xA	1	535258971	537006256	1747285	39	22,32
Wheat4xA	2	15003681	18431007	3427326	103	30,05
Wheat4xA	2	147210723	167699278	20488555	172	8,39
Wheat4xA	2	324337973	336204610	11866637	87	7,33

Wheat4xA	2	390056719	425394515	35337796	247	6,99
Wheat4xA	2	588888627	593154227	4265600	40	9,38
Wheat4xA	2	605602498	624809553	19207055	224	11,66
Wheat4xA	2	637867685	666193736	28326051	332	11,72
Wheat4xA	2	666868043	698285779	31417736	482	15,34
Wheat4xA	2	714696776	743444217	28747441	642	22,33
Wheat4xA	2	752320200	754863195	2542995	57	22,41
Wheat4xA	2	759030981	763115387	4084406	76	18,61
Wheat4xA	3	4792	14292005	14287213	304	21,28
Wheat4xA	3	19644740	24160205	4515465	116	25,69
Wheat4xA	3	53285561	60504926	7219365	129	17,87
Wheat4xA	3	62900860	102574676	39673816	366	9,23
Wheat4xA	3	174673224	184035372	9362148	93	9,93
Wheat4xA	3	467220359	473120990	5900631	36	6,1
Wheat4xA	3	473342649	508180792	34838143	333	9,56
Wheat4xA	3	509236768	524248212	15011444	205	13,66
Wheat4xA	3	634702998	649727941	15024943	267	17,77
Wheat4xA	4	1530554	25124311	23593757	426	18,06
Wheat4xA	4	34156704	37107181	2950477	56	18,98
Wheat4xA	4	62347469	69559138	7211669	60	8,32
Wheat4xA	4	146991341	159520021	12528680	78	6,23
Wheat4xA	4	180554036	189653534	9099498	41	4,51
Wheat4xA	4	491359701	498317446	6957745	83	11,93
Wheat4xA	4	499859995	522179308	22319313	181	8,11
Wheat4xA	4	535558899	540140030	4581131	72	15,72
Wheat4xA	4	549790974	590422226	40631252	725	17,84
Wheat4xA	4	597100649	629593682	32493033	690	21,24
Wheat4xA	4	630583155	671995048	41411893	719	17,36
Wheat4xA	4	675578026	698691169	23113143	374	16,18
Wheat4xA	5	113793647	116754470	2960823	23	7,77
Wheat4xA	5	128367408	141669867	13302459	101	7,59
Wheat4xA	5	419743549	443045405	23301856	303	13
Wheat4xA	5	565007420	570952006	5944586	122	20,52
Wheat4xA	5	651151210	655240949	4089739	69	16,87
Wheat4xA	5	669746151	691465151	21719000	391	18
Wheat4xA	6	33968631	37974245	4005614	82	20,47
Wheat4xA	6	51554292	54889401	3335109	71	21,29
Wheat4xA	6	121192457	146849657	25657200	242	9,43
Wheat4xA	6	156641021	187718213	31077192	228	7,34
Wheat4xA	6	190984929	208018441	17033512	92	5,4
Wheat4xA	6	457526753	466365908	8839155	86	9,73
Wheat4xA	6	533308584	548992070	15683486	163	10,39
Wheat4xA	6	549879563	554749236	4869673	83	17,04
Wheat4xA	6	556885150	564376270	7491120	138	18,42
Wheat4xA	7	192763517	197202371	4438854	50	11,26
Wheat4xA	7	197513182	201658855	4145673	61	14,71
Wheat4xA	7	230261543	238495829	8234286	59	7,17
Wheat4xA	7	265054265	268676210	3621945	25	6,9
Wheat4xA	7	417532615	477739700	60207085	407	6,76
Wheat4xA	7	504547185	511692195	7145010	86	12,04
Wheat4xA	7	563774611	586060626	22286015	217	9,74
Wheat4xA	7	589302371	699905671	110603300	1652	14,94
Wheat4xB	1	327151634	335152166	8000532	77	9,62
Wheat4xB	1	356038678	365238843	9200165	84	9,13
Wheat4xB	1	413330407	429994315	16663908	157	9,42
Wheat4xB	2	13932908	22292162	8359254	188	22,49
Wheat4xB	2	181009177	187739234	6730057	60	8,92

Wheat4xB	2	548664981	569591655	20926674	215	10,27
Wheat4xB	2	584027513	623686204	39658691	392	9,88
Wheat4xB	2	625471499	668084675	42613176	530	12,44
Wheat4xB	2	704499567	755223613	50724046	759	14,96
Wheat4xB	2	785537802	788260425	2722623	61	22,4
Wheat4xB	3	79718834	92525066	12806232	165	12,88
Wheat4xB	3	100224574	147849578	47625004	422	8,86
Wheat4xB	3	153839208	164899142	11059934	114	10,31
Wheat4xB	3	165781932	171319731	5537799	68	12,28
Wheat4xB	3	201521125	207427741	5906616	55	9,31
Wheat4xB	3	452455909	501839570	49383661	438	8,87
Wheat4xB	3	503162764	534509279	31346515	288	9,19
Wheat4xB	3	670035860	689784771	19748911	267	13,52
Wheat4xB	4	8545223	60905135	52359912	835	15,95
Wheat4xB	4	70676489	75946608	5270119	73	13,85
Wheat4xB	4	93553940	116586981	23033041	241	10,46
Wheat4xB	4	411334063	425372596	14038533	115	8,19
Wheat4xB	4	491150285	501186967	10036682	75	7,47
Wheat4xB	4	549834501	602009900	52175399	575	11,02
Wheat4xB	4	602722601	612951769	10229168	119	11,63
Wheat4xB	4	614452292	641356714	26904422	412	15,31
Wheat4xB	5	69866383	76968501	7102118	77	10,84
Wheat4xB	5	387851356	415805665	27954309	329	11,77
Wheat4xB	5	556340087	566201887	9861800	128	12,98
Wheat4xB	6	64100693	74397628	10296935	151	14,66
Wheat4xB	6	189539721	215921669	26381948	216	8,19
Wheat4xB	6	219351537	254401538	35050001	284	8,1
Wheat4xB	6	263982635	273679124	9696489	71	7,32
Wheat4xB	6	473561899	482368677	8806778	79	8,97
Wheat4xB	6	563728713	588922911	25194198	173	6,87
Wheat4xB	6	590353810	601369220	11015410	105	9,53
Wheat4xB	6	606237636	621890271	15652635	184	11,76
Wheat4xB	6	684251596	686777253	2525657	64	25,34
Wheat4xB	6	696575696	703067314	6491618	231	35,58
Wheat4xB	7	159820001	167475103	7655102	97	12,67
Wheat4xB	7	167751045	177233853	9482808	71	7,49
Wheat4xB	7	194595273	213080389	18485116	147	7,95
Wheat4xB	7	281740739	440481197	158740458	939	5,92
Wheat4xB	7	478344286	508683981	30339695	267	8,8
Wheat4xB	7	531021767	551537507	20515740	206	10,04
Wheat4xB	7	553613346	711979696	158366350	2045	12,91
Wheat6xA	1	101320830	104441552	3120722	16	5,13
Wheat6xA	1	293707856	299426015	5718159	33	5,77
Wheat6xA	1	301196754	306120605	4923851	27	5,48
Wheat6xA	1	319077590	325039829	5962239	35	5,87
Wheat6xA	1	333134461	342435218	9300757	60	6,45
Wheat6xA	1	376501862	393459251	16957389	104	6,13
Wheat6xA	1	407330835	420517323	13186488	37	2,81
Wheat6xA	1	536343748	537972802	1629054	38	23,33
Wheat6xA	2	14676530	16875939	2199409	103	46,83
Wheat6xA	2	144608914	161947293	17338379	98	5,65
Wheat6xA	2	391896188	433629639	41733451	133	3,19
Wheat6xA	2	549287799	573379701	24091902	195	8,09
Wheat6xA	2	594994783	599163632	4168849	21	5,04
Wheat6xA	2	611939210	625570739	13631529	140	10,27
Wheat6xA	2	643439927	673328368	29888441	237	7,93
Wheat6xA	2	674024087	704005867	29981780	373	12,44

Wheat6xA	2	721984489	749091641	27107152	462	17,04
Wheat6xA	2	759303759	760600198	1296439	23	17,74
Wheat6xA	2	762640207	780713884	18073677	334	18,48
Wheat6xA	3	20589791	25373886	4784095	96	20,07
Wheat6xA	3	57260777	66225972	8965195	126	14,05
Wheat6xA	3	67148999	107402559	40253560	267	6,63
Wheat6xA	3	107975081	116526742	8551661	78	9,12
Wheat6xA	3	168833301	178261610	9428309	66	7
Wheat6xA	3	359329612	364198495	4868883	25	5,13
Wheat6xA	3	423480751	425607288	2126537	21	9,88
Wheat6xA	3	460147455	465539248	5391793	19	3,52
Wheat6xA	3	466043888	495015975	28972087	208	7,18
Wheat6xA	3	501898052	517080156	15182104	163	10,74
Wheat6xA	3	529106304	544385465	15279161	118	7,72
Wheat6xA	3	556354721	566327988	9973267	78	7,82
Wheat6xA	3	638446092	654071173	15625081	195	12,48
Wheat6xA	4	1725549	24862614	23137065	306	13,23
Wheat6xA	4	34047254	37009594	2962340	36	12,15
Wheat6xA	4	58533307	64571087	6037780	57	9,44
Wheat6xA	4	65456768	75920232	10463464	97	9,27
Wheat6xA	4	452595275	469243601	16648326	114	6,85
Wheat6xA	4	497493612	503617870	6124258	47	7,67
Wheat6xA	4	504789844	514569082	9779238	52	5,32
Wheat6xA	4	543331966	547233472	3901506	42	10,77
Wheat6xA	4	556751829	596828478	40076649	519	12,95
Wheat6xA	4	605640606	639205641	33565035	511	15,22
Wheat6xA	5	94310801	106148506	11837705	47	3,97
Wheat6xA	5	113796642	116582100	2785458	14	5,03
Wheat6xA	5	389052380	391521446	2469066	20	8,1
Wheat6xA	5	430502351	445245106	14742755	151	10,24
Wheat6xA	5	569801814	575667556	5865742	82	13,98
Wheat6xA	5	658670460	662731341	4060881	55	13,54
Wheat6xA	5	677135409	681466878	4331469	76	17,55
Wheat6xA	5	683350382	699470809	16120427	210	13,03
Wheat6xA	6	34979092	39505834	4526742	72	15,91
Wheat6xA	6	53071982	57731978	4659996	51	10,94
Wheat6xA	6	121884386	148251188	26366802	135	5,12
Wheat6xA	6	158046164	187742032	29695868	115	3,87
Wheat6xA	6	412363785	437022553	24658768	99	4,01
Wheat6xA	6	455936783	464690964	8754181	61	6,97
Wheat6xA	6	533256698	549038681	15781983	122	7,73
Wheat6xA	6	550079398	554792868	4713470	57	12,09
Wheat6xA	6	556526686	564886300	8359614	116	13,88
Wheat6xA	7	185981222	190471666	4490444	30	6,68
Wheat6xA	7	194791031	198785324	3994293	30	7,51
Wheat6xA	7	199004637	206005216	7000579	55	7,86
Wheat6xA	7	232742244	240089377	7347133	39	5,31
Wheat6xA	7	266977182	270756258	3779076	17	4,5
Wheat6xA	7	420837199	481251192	60413993	280	4,63
Wheat6xA	7	507819322	515035574	7216252	64	8,87
Wheat6xA	7	564345483	589140409	24794926	168	6,78
Wheat6xA	7	589756475	705691124	115934649	1169	10,08
Wheat6xB	1	149601252	156594405	6993153	44	6,29
Wheat6xB	1	158516376	162814929	4298553	23	5,35
Wheat6xB	1	326763409	331424018	4660609	49	10,51
Wheat6xB	1	349595182	358843009	9247827	55	5,95
Wheat6xB	1	407296070	419772758	12476688	73	5,85

Wheat6xB	2	18264353	26586784	8322431	145	17,42
Wheat6xB	2	548951741	567586269	18634528	143	7,67
Wheat6xB	2	585891499	627074499	41183000	247	6
Wheat6xB	2	628805114	674723350	45918236	406	8,84
Wheat6xB	2	708982118	756606224	47624106	510	10,71
Wheat6xB	3	8024553	10957184	2932631	65	22,16
Wheat6xB	3	22860213	24251194	1390981	25	17,97
Wheat6xB	3	71393282	84764161	13370879	124	9,27
Wheat6xB	3	92604590	140854587	48249997	268	5,55
Wheat6xB	3	146181327	157913743	11732416	75	6,39
Wheat6xB	3	158218255	163709213	5490958	39	7,1
Wheat6xB	3	367835721	373878257	6042536	36	5,96
Wheat6xB	3	421612497	425387270	3774773	28	7,42
Wheat6xB	3	443368058	486383143	43015085	278	6,46
Wheat6xB	3	491519165	523563735	32044570	177	5,52
Wheat6xB	3	659275209	679647705	20372496	181	8,88
Wheat6xB	4	9514869	62965937	53451068	559	10,46
Wheat6xB	4	73298329	78614617	5316288	42	7,9
Wheat6xB	4	95858308	119354687	23496379	153	6,51
Wheat6xB	4	398233713	400695906	2462193	14	5,69
Wheat6xB	4	484769949	495025442	10255493	37	3,61
Wheat6xB	4	543863245	596374763	52511518	383	7,29
Wheat6xB	4	597031988	607291117	10259129	103	10,04
Wheat6xB	4	608623469	636783655	28160186	250	8,88
Wheat6xB	4	660655365	673474837	12819472	193	15,06
Wheat6xB	5	64253267	70913796	6660529	51	7,66
Wheat6xB	5	375168094	405596533	30428439	231	7,59
Wheat6xB	5	550557796	560001598	9443802	95	10,06
Wheat6xB	5	693002831	701462822	8459991	129	15,25
Wheat6xB	6	64230368	75035714	10805346	91	8,42
Wheat6xB	6	92461883	96404580	3942697	40	10,15
Wheat6xB	6	185657380	211164363	25506983	128	5,02
Wheat6xB	6	216468334	255271408	38803074	169	4,36
Wheat6xB	6	259877734	278542581	18664847	43	2,3
Wheat6xB	6	479077207	515578925	36501718	199	5,45
Wheat6xB	6	579355698	602849438	23493740	125	5,32
Wheat6xB	6	604212014	614133168	9921154	62	6,25
Wheat6xB	6	618936336	636397024	17460688	116	6,64
Wheat6xB	7	178418378	196913145	18494767	87	4,7
Wheat6xB	7	334798278	368156975	33358697	67	2,01
Wheat6xB	7	369410910	432503871	63092961	283	4,49
Wheat6xB	7	475663576	504838120	29174544	192	6,58
Wheat6xB	7	526864830	546023385	19158555	144	7,52
Wheat6xB	7	549349552	705256932	155907380	1338	8,58
Wheat6x	1	233307158	240599656	7292498	59	8,09
Wheat6x	1	249559480	253199522	3640042	31	8,52
Wheat6x	1	301629146	313296239	11667093	91	7,8
Wheat6x	1	391308527	394673155	3364628	46	13,67
Wheat6x	2	12619687	14803101	2183414	97	44,43
Wheat6x	2	164284742	172826517	8541775	42	4,92
Wheat6x	2	190993156	192746898	1753742	12	6,84
Wheat6x	2	470210955	486250602	16039647	171	10,66
Wheat6x	2	500791603	528637621	27846018	219	7,86
Wheat6x	2	529177288	562354417	33177129	381	11,48
Wheat6x	2	587162183	619092706	31930523	498	15,6
Wheat6x	2	637519355	650757581	13238226	290	21,91
Wheat6x	3	6402986	8975091	2572105	76	29,55

Wheat6x	3	15943118	16879596	936478	25	26,7
Wheat6x	3	45752205	57159238	11407033	147	12,89
Wheat6x	3	58111496	91223568	33112072	253	7,64
Wheat6x	3	95244740	107046044	11801304	70	5,93
Wheat6x	3	107440644	116040831	8600187	73	8,49
Wheat6x	3	299590798	323720938	24130140	151	6,26
Wheat6x	3	343449782	376273062	32823280	255	7,77
Wheat6x	3	376918749	397783443	20864694	170	8,15
Wheat6x	3	498263989	518592651	20328662	218	10,72
Wheat6x	4	5523443	43072316	37548873	566	15,07
Wheat6x	4	49784556	53762789	3978233	46	11,56
Wheat6x	4	65939637	70238546	4298909	38	8,84
Wheat6x	4	71291915	84548094	13256179	105	7,92
Wheat6x	4	396432635	400657061	4224426	31	7,34
Wheat6x	4	439655754	473811665	34155911	372	10,89
Wheat6x	4	473854998	479961146	6106148	86	14,08
Wheat6x	4	480813768	496003785	15190017	238	15,67
Wheat6x	5	322954908	342961763	20006855	215	10,75
Wheat6x	5	450632474	457997552	7365078	88	11,95
Wheat6x	5	556423412	559896160	3472748	82	23,61
Wheat6x	6	31751236	35631668	3880432	55	14,17
Wheat6x	6	100566882	116341121	15774239	128	8,11
Wheat6x	6	122730282	143664929	20934647	158	7,55
Wheat6x	6	145759235	155544761	9785526	35	3,58
Wheat6x	6	318345544	326541094	8195550	60	7,32
Wheat6x	6	346098635	355016432	8917797	76	8,52
Wheat6x	6	387883951	402349678	14465727	117	8,09
Wheat6x	6	403319426	409293752	5974326	58	9,71
Wheat6x	6	411610565	422081985	10471420	110	10,5
Wheat6x	7	190129210	194252437	4123227	33	8
Wheat6x	7	220507046	228013496	7506450	37	4,93
Wheat6x	7	374564870	416632363	42067493	294	6,99
Wheat6x	7	452710185	478371603	25661418	180	7,01
Wheat6x	7	498527161	515475821	16948660	148	8,73
Wheat6x	7	517153690	613620581	96466891	1108	11,49

Supplementary Table 3: Partitioning of genomes according to recombination rates. Chromosomes are partitioned according to recombination rates in two compartments, respectively LR (Low Recombination) corresponding to the centromeres and peri-centromeres and HR (High recombination) corresponding to the telomeres. Tables indicates for each chromosome (in columns) its size, the postion of the LR region in Mb (LR_start and LR_stop) its size (LR region size) and its ratio compared to whole chromsome size (Ratio LR/ChroSize).

Species_Chro	ChroSize	LR start (pos. Mb)	LR (pos. Mb)	LR region size (Pb)	Ratio
Barley@1H	558425836	90	270	180000000	0,322
Barley@2H	768043353	150	410	260000000	0,339
Barley@3H	699653359	140	350	210000000	0,3
Barley@4H	647060158	100	350	250000000	0,386
Barley@5H	669923547	60	280	220000000	0,328
Barley@6H	583346162	110	350	240000000	0,411
Barley@7H	657183766	210	410	200000000	0,304

Brachy@01	74827228	33	40	7000000	0,094
Brachy@02	59323543	27	34	7000000	0,118
Brachy@03	59853107	23	28	5000000	0,084
Brachy@04	48559729	15	27	12000000	0,247
Brachy@05	28429713	4	11	7000000	0,246
Maize@01	301409969	120	145	25000000	0,083
Maize@02	237798040	80	130	50000000	0,21
Maize@03	232184249	70	105	35000000	0,151
Maize@04	241961162	50	110	60000000	0,248
Maize@05	217885139	95	130	35000000	0,161
Maize@06	169314060	45	55	10000000	0,059
Maize@07	176810119	40	78	38000000	0,215
Maize@08	175288181	40	56	16000000	0,091
Maize@09	156983126	60	80	20000000	0,127
Maize@10	149531493	20	72	52000000	0,348
Setaria@01	58901785	10	25	15000000	0,255
Setaria@02	40675004	12	25	13000000	0,32
Setaria@03	35962510	28	34	6000000	0,167
Setaria@04	36010204	19	23	4000000	0,111
Setaria@05	47251813	13	23	10000000	0,212
Setaria@06	40381311	13	20	7000000	0,173
Setaria@07	50647449	4	11	7000000	0,138
Setaria@08	49198778	15	20	5000000	0,102
Setaria@09	42130900	22	25	3000000	0,071
Sorghum@01	73833847	33	39	6000000	0,081
Sorghum@02	77927413	30	37	7000000	0,09
Sorghum@03	74439055	32	37	5000000	0,067
Sorghum@04	68008026	27	32	5000000	0,074
Sorghum@05	62328788	35	41	6000000	0,096
Sorghum@06	62194152	18	25	7000000	0,113
Sorghum@07	64307977	29	35	6000000	0,093
Sorghum@08	55459831	19	24	5000000	0,09
Sorghum@09	59622314	26	32	6000000	0,101
Sorghum@10	60976732	27	34	7000000	0,115
Wheat4xA@1	593496313	90	290	200000000	0,337
Wheat4xA@2	775170818	230	490	260000000	0,335
Wheat4xA@3	754268017	150	420	270000000	0,358
Wheat4xA@4	726421266	190	420	230000000	0,317
Wheat4xA@5	700752025	100	380	280000000	0,4
Wheat4xA@6	621402681	120	430	310000000	0,499
Wheat4xA@7	727575506	260	490	230000000	0,316
Wheat4xB@1	690493147	170	280	110000000	0,159
Wheat4xB@2	803317284	260	430	170000000	0,212
Wheat4xB@3	841082355	330	390	60000000	0,071
Wheat4xB@4	673852351	200	420	220000000	0,326
Wheat4xB@5	712160875	120	400	280000000	0,393
Wheat4xB@6	703074765	220	430	210000000	0,299
Wheat4xB@7	755249505	270	400	130000000	0,172
Wheat6xA@1	594002067	80	280	200000000	0,337
Wheat6xA@2	780778433	100	440	340000000	0,435
Wheat6xA@3	750741511	210	450	240000000	0,32
Wheat6xA@4	744525004	150	450	300000000	0,403
Wheat6xA@5	709676514	80	350	270000000	0,38
Wheat6xA@6	617935868	190	400	210000000	0,34
Wheat6xA@7	736698113	240	400	160000000	0,217
Wheat6xB@1	689851782	170	280	110000000	0,159
Wheat6xB@2	801255821	200	390	190000000	0,237

Wheat6xB@3	830611874	220	390	170000000	0,205
Wheat6xB@4	673535148	170	410	240000000	0,356
Wheat6xB@5	713118731	140	240	100000000	0,14
Wheat6xB@6	720954733	280	430	150000000	0,208
Wheat6xB@7	750611715	230	360	130000000	0,173
Wheat6xD@1	495424774	80	190	110000000	0,222
Wheat6xD@2	651824161	160	320	160000000	0,245
Wheat6xD@3	615475387	170	310	140000000	0,227
Wheat6xD@4	509854941	100	300	200000000	0,392
Wheat6xD@5	565977789	70	230	160000000	0,283
Wheat6xD@6	473517451	140	280	140000000	0,296
Wheat6xD@7	638668735	240	370	130000000	0,204

Supplementary Table 4: LF/MF and Singleton/Pair contents in cereal genomes. For each species (in lines) are indicated (in columns) the numbers of genes for LF and MF compartements and for Pairs (Pair) and singletons (Single) defined by the synteny-based and phylogeny-based methods (see main manuscript).

	Fraction		Synteny-based		Phylogeny-based	
Species	LF	MF	Pair	Single	Pair	Single
Barley	14848	5660	900	5366	7828	14587
Brachypodium	12233	7036	2474	7813	8599	12655
Maize	9155	4238	2590	8509	10148	14946
Maize WGD	29148	17455	7139	8898	14293	15157
Rice	16738	9645	2560	7779	10757	12358
Setaria	14486	7555	1954	7256	12212	12333
Sorghum	10925	6105	2444	7608	10070	12622
Wheat4xA	17947	8580	1040	6233	8727	13624
Wheat4xB	19251	8018	992	6075	8817	13938
Wheat6xA	13290	5673	1342	6379	12021	16402
Wheat6xB	14395	5961	1340	6194	12375	16876
Wheat6xD	14856	5840	1396	6450	12283	16219
TOTAL	187272	91766	26171	84560	128130	171717

Supplementary Table 5: Branch lengths values for cereals and Triticeae phylogenetic trees. A phylogenetic tree branch is designated either by the nodes leading to either ancestral or modern genomes-species (in lines). Each branch is characterized by (in column) its duration (time interval in My), its expected substitutions per site value and its substitution rate corresponding to expected substitutions per site value normalized by branch duration in billion years.

Cereals

Branch	**Time interval (mya)**	**Expected substitutions per site**	**Substitution rate**
AGK7 - AGK12	30	0,123152901	4,11
AGK12 - (Rice,Brachypodium)	14	0,030097206	2,15

(Rice)	46	0,125974411	2,74
(Rice,Brachypodium) - (Barley,Brachypodium)	11	0,055405233	5,04
(Brachypodium)	35	0,075830654	2,17
(Brachypodium) - (Barley,Wheats)	22	0,071019049	3,23
(Barley)	13	0,039960207	3,07
(Barley,Wheats) - Wheats root	6,5	0,016169159	2,49
Wheats6xD	6,5	0,014464282	2,23
Wheats root - (Wheat4xA,Wheats6xA)	6,5	0,01842907	2,84
Wheats root - (Wheat4xB,Wheats6xB)	6,5	0,017413561	2,68
AGK12 - (Maize,Setaria)	33	0,067276997	2,04
(Setaria)	27	0,062491641	2,31
(Maize,Setaria) - (Maize,Sorghum)	11	0,047269594	4,30
(Sorghum)	16	0,030906838	1,93
(Maize)	16	0,059734664	3,73
(Maize preWGD)	11	0,01079388	0,98
(Maize PostWGD)	5	0,048940784	9,79

Triticeae

Branch	**Time interval (Mya)**	**Expected substitutions per site**	**substitution rate**
Barley	13	0,0383771	2,95
Wheat post Barley divergence before ABD divergence	6,5	0,0127541	1,96
WheatA,T. urtatu ancestor	6,04	0,01885704	3,12
T. urartu	0,46	0,00345117	7,50
WheatA post T. urartu divergence before tetraploidy	0,1	0,00180256	18,03
Wheat4xA	0,36	0,00496338	13,79
Wheat6xA	0,36	0,00243867	6,77
WheatB post ABD divergence before tetraploidy	6,14	0,0135837	2,21
Wheat4xB	0,36	0,00554989	15,42
Wheat6xB	0,36	0,0027314	7,59
WheatD post ABD divergence before A. tauschii divergence	6,2	0,01636904	2,64
A. tauschii	0,3	0,00489729	16,32
Wheat6xD (cumulate diplo/hexaplo)	0,3	0,00238688	7,96

Supplementary Table 6: Substitution rate and branch length in cereals for LF/MF and Singleton/Pair genes. Ka, Ks, Ka/Ks and branch length for LF-MF and Singleton/Pair genes (in lines) are described for the inferred ancestors. Notes: "Ks pair/single" corresponds to the Ks ratio of genes in pairs out of singletons; non-significant (NS, in red); significance (in green) is declared according to a 1% threshold on a two-sided bootstrap interval.

Sig. Threshold:1% 2-sided	AGK12	SMS	MS	RBWB	BWB	BW	WABD	WA	WB
Ka pair/single	0.982	NS	1.100	NS	0.935	NS	NS	NS	NS
Ks pair/single	1.006	NS	1.072	1.024	1.059	1.149	1.089	NS	NS
KaKs pair/single	NS	1.154	1.135	NS	1.111	NS	0.947	NS	2.301
Branch length pair/single	0.879	1.034	1.101	0.944	0.965	1.067	1.054	NS	NS
Ka MF/LF	1.046	1.102	1.091	1.026	NS	NS	NS	1.189	NS
Ks MF/LF	1.017	1.033	1.072	1.027	1.028	NS	NS	NS	NS
KaKs MF/LF	0.94	1.109	NS	NS	1.059	NS	NS	1.106	0.856
Branch length MF/LF	1.019	1.133	1.112	1.034	1.058	1.058	NS	1.177	NS

Supplementary Table 7: Global DNA methylation in Brachypodium and maize. Percentage of methylation of the cytosines in the CG, CGH, CHH contexts for maize and Brachypodium for three developmental stages (expressed in days after anthesis) corresponding to wheat developmental stages expressed in degrees day (DD) with two replicates per stage and per species.

	Sample IDs	Stage (Day after anthesis)	CG (%)	CHG (%)	CHH (%)	Corresponding stage In Wheat
Brachy	2-A1	9	64.73	39.54	8.81	100DD
	2-A2	9	63.62	38.25	8.16	100DD
	2-A3	16	66.27	38.5	7.39	250DD
	2-A4	16	59.85	34.5	7.18	250DD
	2-A5	28	72.04	45.64	14.78	500DD
	2-A6	28	66.56	39.91	12.95	500DD
Maize	Sample 7	7	80.44	68.37	3.120	100DD
	Sample 8	7	80.21	68.39	3.245	100DD
	Sample 9	15	80.10	64.74	3.086	250DD
	Sample 10	15	80.23	63.97	3.124	250DD
	Sample 11	35	81.27	64.23	4.257	500DD
	Sample 12	35	81.38	64.74	4.269	500DD

<u>Supplementary Table 8:</u> Comparison of omic features between wheat subgenomes (A, B, D and LF, MF). Multiple comparison using the Tuckey method of means values for (in lines) expression, SNPs density, Ka and Ks between hexaploid wheat A, B and D subgenomes and LF and MF compartments (second column). Notes: diff: difference between means of the two groups; lwr and upr: the lower and the upper end point of the confidence interval at 95%; p adj: p-value after adjustment for the multiple comparisons.

	MF/LF	diff	lwr	upr	p.adj	conclusion		
Expression	MF:Wheat6xA-LF:Wheat6xA	0.028531523	-0.017505056	0.0745681030	0.4878528	ns	?	**MF-D=MF-A=LF-A>MF-B>LF-B<LF-D**
	MF:Wheat6xB-LF:Wheat6xB	0.046035901	0.001371798	0.0907000042	0.0388817	MF-B>LF-B		
	MF:Wheat6xD-LF:Wheat6xD	0.078197497	0.033484201	0.1229107928	0.0000092	MF-D>LF-D		
	LF:Wheat6xB-LF:Wheat6xA	-0.078392809	-0.113303696	-0.0434819220	0.0000000	LF-B<LF-A	LF-A>LF-D>LF-B	
	LF:Wheat6xD-LF:Wheat6xA	-0.035539890	-0.070161006	-0.0009187736	0.0402565	LF-D<LF-A		
	LF:Wheat6xD-LF:Wheat6xB	0.042852919	0.008938936	0.0767669017	0.0042901	LF-D>LF-B		
	MF:Wheat6xB-MF:Wheat6xA	-0.060888431	-0.114698084	-0.0070787785	0.0159473	MF-B<MF-A	MF-D=MF-A>MF-B	
	MF:Wheat6xD-MF:Wheat6xA	0.014126084	-0.039911161	0.0681633282	0.9762576	ns		
	MF:Wheat6xD-MF:Wheat6xB	0.075014515	0.021685323	0.1283437069	0.0008654	MF-D>MF-B		
SNPs	MF:Wheat6xA-LF:Wheat6xA	0.001991612	-0.004661149	0.008644372	0.9573371	ns	?	**MF-B>LF-B>MF-A=LF-A>MF-D=LF-D**
	MF:Wheat6xB-LF:Wheat6xB	0.015755360	0.009153461	0.022357258	0.0000000	MF-B>LF-B		
	MF:Wheat6xD-LF:Wheat6xD	0.004293253	-0.002265127	0.010851633	0.4236407	ns		
	LF:Wheat6xB-LF:Wheat6xA	0.017160234	0.012250245	0.022070224	0.0000000	LF-B>LF-A	LF-B>LF-A>LF-D	
	LF:Wheat6xD-LF:Wheat6xA	-0.047762011	-0.052630227	-0.042893796	0.0000000	LF-D<LF-A		
	LF:Wheat6xD-LF:Wheat6xB	-0.064922246	-0.069772869	-0.060071622	0.0000000	LF-D<LF-B		
	MF:Wheat6xB-MF:Wheat6xA	0.030923982	0.022940482	0.038907483	0.0000000	MF-B>MF-A	MF-B>MF-A>MF-D	
	MF:Wheat6xD-MF:Wheat6xA	-0.045460370	-0.053433578	-0.037487162	0.0000000	MF-D<MF-A		
	MF:Wheat6xD-MF:Wheat6xB	-0.076384353	-0.084325943	-0.068442762	0.0000000	MF-D<MF-B		
ka	MF:Wheat6xA-LF:Wheat6xA	2.744009e-03	-0.0012883693	0.006776388	0.3780524	ns	MF=LF	**MF-D=MF-A=MF-B=LF-A=LF-D=LF-B**
	MF:Wheat6xB-LF:Wheat6xB	1.921163e-03	-0.0020802782	0.005922603	0.7461866	ns		
	MF:Wheat6xD-LF:Wheat6xD	3.558244e-03	-0.0004160536	0.007532541	0.1095743	ns		
	LF:Wheat6xB-LF:Wheat6xA	-2.119962e-04	-0.0031869713	0.002762979	0.9999527	ns	LF-A=LF-D=LF-B	
	LF:Wheat6xD-LF:Wheat6xA	-2.591297e-04	-0.0032090107	0.002690751	0.9998671	ns		
	LF:Wheat6xD-LF:Wheat6xB	-4.713355e-05	-0.0029861862	0.002891919	1.0000000	ns		
	MF:Wheat6xB-MF:Wheat6xA	-1.034843e-03	-0.0058743811	0.003804695	0.9904034	ns	MF-D=MF-A=MF-B	
	MF:Wheat6xD-MF:Wheat6xA	5.551046e-04	-0.0042774225	0.005387632	0.9995040	ns		
	MF:Wheat6xD-MF:Wheat6xB	1.589948e-03	-0.0032234229	0.006403318	0.9357829	ns		
ks	MF:Wheat6xA-LF:Wheat6xA	0.0170721748	0.009130219	0.025014131	0.0000000	MF-A>LF-A	MF>LF	**MF-D=MF-A=MF-B>LF-A=LF-D=LF-B**
	MF:Wheat6xB-LF:Wheat6xB	0.0114384807	0.003556965	0.019319996	0.0005060	MF-B>LF-B		
	MF:Wheat6xD-LF:Wheat6xD	0.0160235295	0.008196048	0.023851012	0.0000001	MF-D>LF-D		
	LF:Wheat6xB-LF:Wheat6xA	0.0030612515	-0.002798989	0.008921493	0.6716309	ns	LF-A=LF-D=LF-B	
	LF:Wheat6xD-LF:Wheat6xA	-0.0007005632	-0.006510625	0.005109499	0.9993717	ns		
	LF:Wheat6xD-LF:Wheat6xB	-0.0037618147	-0.009551260	0.002027631	0.4324052	ns		
	MF:Wheat6xB-MF:Wheat6xA	-0.0025724425	-0.012103997	0.006959112	0.9726900	ns	MF-D=MF-A=MF-B	
	MF:Wheat6xD-MF:Wheat6xA	-0.0017492084	-0.011266946	0.007768529	0.9952555	ns		
	MF:Wheat6xD-MF:Wheat6xB	0.0008232341	-0.008656749	0.010303217	0.9998744	ns		

Supplementary Table 9: Expression and methylation profiles of conserved genes. Expression and methylation modules were defining using thresholds of 1 TPM for expression and 5 CpG for methylation. Namely expressed genes have an expression level above 1 TPM and methylated genes have at least 5 CpG. Genes above the thresholds are recorded as expressed/methylated (noted as 1) and genes below the threshold are not expressed/methylated (noted as 0) for each of the three developmental stages. Notes: 1-1-1 corresponds to an expression/methylation in all the three stages; 0-0-0: genes not expressed/methylated in the three stages; On-Off: genes with inverse expression/methylation patterns; else: Expression/methylation differences between the three developmental stages.

		1-1-1	0-0-0	1-0 or 0-1	else	total
Expression	Brachy	1475 (74%)	424 (21%)	X	98 (5%)	1997
	Maize 1|2	1156 (58%)	248 (12%)	249 (13%)	344 (17%)	1997
	Wheat A|B	1273 (64%)	285 (14%)	30 (2%)	409 (20%)	1997
	Wheat B|D	1285 (64%)	286 (14%)	25 (1%)	401 (20%)	1997
	Wheat A|D	1288 (64%)	277 (14%)	30 (2%)	402 (20%)	1997
	Wheat A|B|D	1233 (62%)	258 (13%)	X	506 (25%)	1997
	Brachy-Maize	1006 (50%)	143 (7%)	209 (10%)	639 (32%)	1997
	Brachy-Wheat	1094 (55%)	170 (9%)	180 (9%)	553 (27%)	1997
	Maize-Wheat	977 (49%)	157 (8%)	24 (1%)	839 (42%)	1997
	Brachy-Maize-Wheat	868 (43%)	116 (6%)	X	1013 (51%)	1997
Methylation	Brachy	143 (5%)	1454 (49%)	X	1356 (46%)	2953
	Maize 1|2	657 (22%)	676 (23%)	609 (21%)	1011 (34%)	2953
	Brachy-Maize 1	80 (3%)	702 (24%)	514 (17%)	1657 (56%)	2953
	Brachy-Maize 2	87 (3%)	673 (23%)	531 (18%)	1662 (56%)	2953
	Brachy-Maize 1|2	59 (2%)	410 (14%)	249 (8%)	2235 (76%)	2953

3 Références de l'article

Acharya, D. and Ghosh, T.C. (2016) Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics*, **17**, 1–14.

Alekseyev, M.A. and Pevzner, P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, **19**, 943–957.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.

Aury, J., Jaillon, O., Duret, L. and Al., E. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**, 171–178.

Bewick, A.J. and Schmitz, R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.*, **36**, 103–110.

Birchler, J.A. and Veitia, R.A. (2011) Protein-protein and protein-DNA dosage balance and differential paralog transcription factor retention in polyploids. *Front. Plant Sci.*, **2**, 1–3.

Birchler, J.A. and Veitia, R.A. (2014) The gene balance hypothesis: Dosage effects in plants. *Methods Mol. Biol.*, **1112**, 25–32.

Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D. and Veitia, R.A. (2010) Heterosis. *Plant Cell*, **22**, 2105–2112.

Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.*, **13**, 137–144.

Buggs, R.J.A., Elliott, N.M., Zhang, L., Koh, J., Viccini, L.F., Soltis, D.E. and Soltis, P.S. (2010) Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *New Phytol.*, **186**, 175–183.

Bukowski, R., Guo, X., Lu, Y., et al. (2018) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, 1–12.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Chalhoub, B., Denoeud, F., Liu, S., et al. (2014) Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. *Science*, **345**, 950–953.

Chang, F., Gu, Y., Ma, H. and Yang, Z. (2013) AtPRK2 promotes ROP1 activation *via* RopGEFs in the control of polarized pollen tube growth. *Mol. Plant*, **6**, 1187–1201.

Chapman, B.A., Bowers, J.E., Feltus, F.A. and Paterson, A.H. (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2730–2735.

Charif, D. and Lobry, J.R. (2007) SeqinR 1 . 0-2 : A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis Seqin R 1 . 0-

2 : a contributed package to the R project for statistical computing devoted to biological sequences ret. In *Structural approaches to sequence evolution: Molecules, networks, populations*. pp. 207–232.

Chen, Y., Pal, B., Visvader, J.E., Smyth, G.K., Andrews, S. and Macdonald, J.W. (2018) Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. *F1000Research*, 1–40.

Cheng, F., Sun, R., Hou, X., et al. (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nat. Genet.*, **48**, 1218–1224.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G. and Wang, X. (2012) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One*, **7**.

Cheng, X.F., Wittich, P.E., Kieft, H., Angenent, G., XuHan, X. and Lammeren, A.A.M. Van (2000) Temporal and spatial expression of MADS box genes, FBP7 and FBP11, during initiation and early development of ovules in wild type and mutant Petunia hybrida. *Plant Biol.*, **2**, 693–702.

Chuang, T.J., Chen, F.C. and Chen, Y.Z. (2012) Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 15841–15846.

Chueasiri, C., Chunthong, K., Pitnjam, K., et al. (2014) Rice ORMDL controls sphingolipid homeostasis affecting fertility resulting from abnormal pollen development. *PLoS One*, **9**.

Coate, J.E., Farmer, A.D., Schiefelbein, J.W. and Doyle, J.J. (2020) Expression Partitioning of Duplicate Genes at Single Cell Resolution in Arabidopsis Roots. *Front. Genet.*, **11**.

Colombo, L., Franken, J., Krol, A.R. Van Der, Wittich, P.E., Dons, H.J.M. and Angenent, G.C. (1997) Downregulation of ovule-specific MADS box genes from petunia results in maternally controlled defects in seed development. *Plant Cell*, **9**, 703–715.

Conant, G.C., Birchler, J.A. and Pires, J.C. (2014) Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.*, **19**, 91–98.

Darwin, C.R. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* J. Murray, ed., London.

Davis, M.P.A., Dongen, S. Van, Abreu-goodger, C., Bartonicek, N. and Enright, A.J. (2023) Kraken : A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.

Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**.

DeSmet, R., Adams, K.L., Vandepoele, K., Montagu, M.C.E. Van, Maere, S. and Peer, Y. Van de (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci.*, **110**, 2898–2903.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and DePamphilis, C.W. (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol. Biol. Evol.*, **23**, 469–478.

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C.M., Deal, K.R., Dai, X. and Dawson, M.W. (2018) Structural variation and rates of genome evolution in the grass family seen through comparison of sequences of genomes greatly differing in size. *Plant J.*, **95**, 487–503.

Edgar, R.C. (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 1–19.

Edger, P.P., Poorten, T.J., Vanburen, R., et al. (2019) Origin and evolution of the octoploid strawberry genome. *Nat. Genet.*, **51**, 541–547.

Edger, P.P., Smith, R., McKain, M.R., et al. (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell*, **29**, 2150–2167.

El-Sharkawy, I., Mila, I., Bouzayen, M. and Jayasankar, S. (2010) Regulation of two germin-like protein genes during plum fruit development. *J. Exp. Bot.*, **61**, 1761–1770.

Escudero, M. and Wendel, J.F. (2020) The grand sweep of chromosomal evolution in angiosperms. *New Phytol.*, **228**, 805–808.

Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G., Yanofsky, M.F., Kater, M.M. and Colombo, L. (2003) MADS-Box Protein Complexes Control Carpel and Ovule Development in Arabidopsis. *Plant Cell*, **15**, 2603–2611.

Folta, K.M. and Barbey, C.R. (2019) The strawberry genome: a complicated past and promising future. *Hortic. Res.*, **6**, 19–21.

Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L. and Peer, Y. Van de (2020) Polyploidy: A Biological Force From Cells to Ecosystems. *Trends Cell Biol.*, **30**, 688–694.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.*, **60**, 433–453.

Freeling, M., Scanlon, M.J. and Fowler, J.F. (2015) Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.*, **35**, 110–118.

Glastad, K.M., Gokhale, K., Liebig, J. and Goodisman, M.A.D. (2016) The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Sci. Rep.*, **6**, 1–14.

Glémin, S., Scornavacca, C., Dainat, J., et al. (2019) Pervasive hybridizations in the history of wheat relatives. *Sci. Adv.*, **5**, 1–10.

Goidts, V., Szamalek, J.M., Jong, P.J. De, Cooper, D.N., Chuzhanova, N., Hameister, H. and Kehrer-Sawatzki, H. (2005) Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res.*, **15**, 1232–1242.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., et al. (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.*, **8**.

Gu, Y.Q., Anderson, O.D., Londeorë, C.F., Kong, X., Chibbar, R.N. and Lazo, G.R. (2003) Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes. *Genome*, **46**, 1084–1097.

Hancock, J.F. (2005) Contributions of domesticated plant studies to our understanding of plant evolution. *Ann. Bot.*, **96**, 953–963.

Hannweg, K., Steyn, W. and Bertling, I. (2016) In vitro-induced tetraploids of Plectranthus esculentus are nematode-tolerant and have enhanced nutritional value. *Euphytica*, **207**, 343–351.

Hao, Yue, Mabry, M.E., Edger, P.P., et al. (2021) The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. *Genome Res.*, **31**, 799–810.

Hao, Yaqi, Zong, X., Ren, P., Qian, Y. and Fu, A. (2021) Basic helix-loop-helix (Bhlh) transcription factors regulate a wide range of functions in arabidopsis. *Int. J. Mol. Sci.*, **22**.

Hias, N., Svara, A. and Keulemans, J.W. (2018) Effect of polyploidisation on the response of apple (Malus × domestica Borkh.) to Venturia inaequalis infection. *Eur. J. Plant Pathol.*, **151**, 515–526.

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R. and Leitch, I.J. (2017) Is There an Upper Limit to Genome Size? *Trends Plant Sci.*, **22**, 567–573.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A. and Haeseler, A. Von (2018) MPBoot : fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.*, **18**, 1–11.

Ilyas, M., Rasheed, A. and Mahmood, T. (2016) Functional characterization of germin and germin-like protein genes in various plant species using transgenic approaches. *Biotechnol. Lett.*, **38**, 1405–1421.

Joron, M., Frezal, L., Jones, R.T., et al. (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C. and Soltis, P.S. (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.*, **105**, 348–363.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004) Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics*, **166**, 935–945.

LeComber, S. and Smith, C. (2004) Polyploidy in fishes: patterns and processes. *Biol. J. Linn. Soc.*, **82**, 431–442.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., et al. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.

Leitch, A.R. and Leitch, I.J. (2008) Genomic plasticity and the diversity of polyploid plants. *Science*, **320**, 481–483.

Leitch, I.J. and Bennett, M.D. (1997) Polyploidy in angiosperms. *Trends Plant Sci.*, **2**, 470–476.

Levin, D.A. and Soltis, D.E. (2018) ScienceDirect Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K – Pg interlude. *Curr. Opin. Plant Biol.*, **42**, 1–7.

Li, B. and Dewey, C.N. (2011) RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 1–16.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics*, **00**, 1–3.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M. and Muehlbauer, G.J. (2016) Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC Genomics*, **17**, 875.

Li, M., Wang, R., Wu, X. and Wang, J. (2020) Homoeolog expression bias and expression level dominance (ELD) in four tissues of natural allotetraploid Brassica napus. *BMC Genomics*, **21**, 1–15.

Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K., Li, L. and Zhang, S. (2019) Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). *Hortic. Res.*, **6**, 1–12.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H. and Barker, M.S. (2015) Early genome duplications in conifers and other seed plants. *Sci. Adv.*, **1**, 1–7.

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J. and Barker, M.S. (2018) Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 4713–4718.

Lian, S., Zhou, Y., Liu, Z., Gong, A. and Cheng, L. (2020) The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences. *BMC Plant Biol.*, **20**, 1–16.

Liao, B.Y. and Zhang, J. (2006) Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol. Biol. Evol.*, **23**, 1119–1128.

Lim, C.W., Baek, W., Han, S.W. and Lee, S.C. (2013) Arabidopsis PYl8 plays an important role for ABA signaling and drought stress responses. *Plant Pathol. J.*, **29**, 471–476.

Liu, S., Liu, Y., Yang, X., et al. (2014) The brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.*, **5**, 1–11.

Liu, X.L., Jiang, F.F., Wang, Z.W., et al. (2017) Wider geographic distribution and higher diversity of hexaploids than tetraploids in Carassius species complex reveal recurrent polyploidy

effects on adaptive evolution. *Sci. Rep.*, **7**, 1–10.

Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N., Sun, S., Zhang, L. and Wang, X. (2017) Two highly similar poplar paleo-subgenomes suggest an autotetraploid ancestor of salicaceae plants. *Front. Plant Sci.*, **8**, 1–11.

Lovell, J.T., MacQueen, A.H., Mamidi, S., et al. (2021) Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, **590**, 438–444.

Lu, K., Wei, L., Li, X., et al. (2019) Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat. Commun.*, **10**, 1–12.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Ma, Y.Q., Pu, Z.Q., Tan, X.M., Meng, Q., Zhang, K.L., Yang, L., Ma, Y.Y., Huang, X. and Xu, Z.Q. (2022) SEPALLATA-like genes of Isatis indigotica can affect the architecture of the inflorescences and the development of the floral organs. *PeerJ*, **10:e13034**, 1–25.

Marais, D.L. Des and Rausher, M.D. (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, **454**, 762–765.

Maria Maggiolini, F.A., Sanders, A.D., Shew, C.J., et al. (2020) Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.*, **30**, 1680–1693.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 10–12.

Mccarthy, D.J., Chen, Y. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D. and Duret, L. (2012) High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics*, **28**, 1078–1085.

Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., et al. (2022) Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, **602**, 101–105.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. and Salse, J. (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.*, **16**, 262.

Murat, F., Zhang, R., Guizard, S., et al. (2014) Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biol. Evol.*, **6**, 12–33.

Murphy, W.J., Larkin, D.M., Everts-Van Der Wind, A., et al. (2005) Evolution: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*,

309, 613–617.

Naser-Khdour, S., Quang Minh, B., Zhang, W., Stone, E.A. and Lanfear, R. (2019) The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biol. Evol.*, **11**, 3341–3352.

Nguyen, L., Schmidt, H.A., Haeseler, A. Von and Minh, B.Q. (2014) IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag.

Page, J.T., Liechty, Z.S., Alexander, R.H., Clemons, K., Hulse-Kemp, A.M., Ashrafi, H., Deynze, A. Van, Stelly, D.M. and Udall, J.A. (2016) DNA Sequence Evolution and Rare Homoeologous Conversion in Tetraploid Cotton. *PLoS Genet.*, **12**, 1–22.

Parkin, I.A.P., Koh, C., Tang, H., et al. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol.*, **15**, 1–18.

Parks, M.B., Nakov, T., Ruck, E.C., Wickett, N.J. and Alverson, A.J. (2018) Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *Am. J. Bot.*, **105**, 330–347.

Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Paterson, A.H., Wendel, J.F., Gundlach, H., et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, **492**, 423–427.

Peer, Y. Van De, Mizrachi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.*, **18**, 411–424.

Pham, S.K. and Pevzner, P.A. (2010) DRIMM-Synteny: Decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.

Pons, P. and Latapy, M. (2006) Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, **10**, 191–218.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-korts, D. and Goué, N. (2019) Tracing the ancestry of modern bread wheats. *Nat. Genet.*, **51**, 905–911.

Pont, C., Murat, F., Guizard, S., et al. (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant J.*, **76**, 1030–1044.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C. and Salse, J. (2019) Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.*, **20**, 1–17.

Ramírez-González, R.H., Borrill, P., Lang, D., et al. (2018) The transcriptional landscape of polyploid wheat. *Science*, **361**, 1–12.

Renny-Byfield, S., Gallagher, J.P., Grover, C.E., Szadkowski, E., Page, J.T., Udall, J.A., Wang, X., Paterson, A.H. and Wendel, J.F. (2014) Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.*, **6**, 559–571.

Renny-Byfield, S., Gong, L., Gallagher, J.P. and Wendel, J.F. (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol. Biol. Evol.*, **32**, 1063–1071.

Renny-Byfield, S., Rodgers-Melnick, E. and Ross-Ibarra, J. (2017) Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol.*, **34**, 1825–1832.

Renny-Byfield, S. and Wendel, J.F. (2014) Doubling down on genomes: Polyploidy and crop plants. *Am. J. Bot.*, **101**, 1711–1725.

Robinson, M.D., Mccarthy, D.J. and Smyth, G.K. (2010) edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Salman-Minkov, A., Sabath, N. and Mayrose, I. (2016) Whole-genome duplication as a key factor in crop domestication. *Nat. Plants*, **2**, 1–4.

Salse, J. (2016) Deciphering the evolutionary interplay between subgenomes following polyploidy: A paleogenomics approach in grasses. *Am. J. Bot.*, **103**, 1167–1174.

Salse, J. (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.*, **15**, 122–130.

Sarda, S., Zeng, J., Hunt, B.G. and Yi, S. V. (2012) The evolution of invertebrate gene body methylation. *Mol. Biol. Evol.*, **29**, 1907–1916.

Schmid, M., Evans, B.J. and Bogart, J.P. (2015) Polyploidy in Amphibia. *Cytogenet. Genome Res.*, **145**, 315–330.

Schmutz, J., Cannon, S.B., Schlueter, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.*, **4**, 265–277.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 4069–4074.

Scornavacca, C., Jacox, E. and Szöllosi, G.J. (2015) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31**, 841–848.

Seoighe, C. and Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.*, **2**, 548–554.

Shi, J., Ma, X., Zhang, J., et al. (2019) Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.*, **10**, 1–9.

Shi, T., Rahmani, R.S., Gugger, P.F., et al. (2020) Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering

Plants. *Mol. Biol. Evol.*, **37**, 2394–2413.

Slane, D., Reichardt, I., Kasmi, F. El, Bayer, M. and Jürgens, G. (2017) Evolutionarily diverse SYP1 Qa-SNAREs jointly sustain pollen tube growth in Arabidopsis. *Plant J.*, **92**, 375–385.

Soltis, P.S. and Soltis, D.E. (2009) The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.*, **60**, 561–588.

Song, K., Li, L. and Zhang, G. (2018) Relationship Among Intron Length, Gene Expression, and Nucleotide Diversity in the Pacific Oyster Crassostrea gigas. *Mar. Biotechnol.*, **20**, 676–684.

Stamatakis, A. (2014) RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S. V., Obermeier, C., Parkin, I.A.P., Chèvre, A.M. and Snowdon, R.J. (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus. *Plant Biotechnol. J.*, **15**, 1478–1489.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., et al. (2007) Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics*, **176**, 2055–2067.

Sun, H., Wu, S., Zhang, G., et al. (2017) Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Mol. Plant*, **10**, 1293–1306.

Sun, Y., Zhong, M., Li, Y., Zhang, R., Su, L., Xia, G. and Wang, H. (2021) GhADF6-mediated actin reorganization is associated with defence against Verticillium dahliae infection in cotton. *Mol. Plant Pathol.*, **22**, 1656–1667.

Suyama, M., Torrents, D., Bork, P. and Delbru, M. (2006) PAL2NAL : robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, 609–612.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

Szöllosi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E. and Daubin, V. (2013) Efficient exploration of the space of reconciled gene trees. *Syst. Biol.*, **62**, 901–912.

Takuno, S. and Gaut, B.S. (2012) Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. *Mol. Biol. Evol.*, **29**, 219–227.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A. and Peer, Y. Van de (2003) Genome duplication, a trait shared by 22 000 species of ray-finned fish. *Genome Res.*, **13**, 382–390.

Thomas, B.C., Pedersen, B. and Freeling, M. (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**, 934–946.

Throude, M., Bolot, S., Bosio, M., et al. (2009) Structure and expression analysis of rice paleo duplications. *Nucleic Acids Res.*, **37**, 1248–1259.

Tsunoyama, K., Bellgard, M.I. and Gojobori, T. (2001) Intragenic variation of synonymous

substitution rates is caused by nonrandom mutations at methylated CpG. *J. Mol. Evol.*, **53**, 456–464.

Ulrich, D. and Olbricht, K. (2013) Diversity of volatile patterns in sixteen Fragaria vesca L. Accessions in comparison to cultivars of Fragaria xananassa. *J. Appl. Bot. Food Qual.*, **86**, 37–46.

VanBuren, R., Man Wai, C., Wang, X., et al. (2020) Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.*, **11**, 1–11.

Vanneste, K., Baele, G., Maere, S. and Peer, Y. Van De (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous − Paleogene boundary. *Genome Res.*, **32**, 1334–1347.

Wang, M., Tu, L., Lin, M., et al. (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.*, **49**, 579–587.

Wang, X., Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L., Wendel, J.F. and Liu, B. (2017) Gene-body CG methylation and divergent expression of duplicate genes in rice. *Sci. Rep.*, **7**, 1–11.

Wang, Y., Wang, X., Lee, T.H., Mansoor, S. and Paterson, A.H. (2013) Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). *New Phytol.*, **198**, 274–283.

Wang, Y., Wang, X. and Paterson, A.H. (2012) Genome and gene duplications and gene expression divergence: A view from plants. *Ann. N. Y. Acad. Sci.*, **1256**, 1–14.

Wickland, D.P. and Hanzawa, Y. (2015) The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. *Mol. Plant*, **8**, 983–997.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. and Rieseberg, L.H. (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 13875–13879.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M. and Wang, X. (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci.*, **111**, 5283–5288.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. and Freeling, M. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.*, **8**.

Xu, W., Zhang, Q., Yuan, W., et al. (2020) The genome evolution and low-phosphorus adaptation in white lupin. *Nat. Commun.*, **11**, 1–13.

Yang, L. and Gaut, B.S. (2011) Factors that contribute to variation in evolutionary rate among arabidopsis genes. *Mol. Biol. Evol.*, **28**, 2359–2369.

Yim, W.C., Lee, B.M. and Jang, C.S. (2009) Expression diversity and evolutionary dynamics of rice duplicate genes. *Mol. Genet. Genomics*, **281**, 483–493.

Zhang, J., Liu, Y., Xia, E.H., Yao, Q.Y., Liu, X.D. and Gao, L.Z. (2015) Autotetraploid rice

methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E7022–E7029.

Zhang, Q., Guan, P., Zhao, L., et al. (2021) Asymmetric epigenome maps of subgenomes reveal imbalanced transcription and distinct evolutionary trends in Brassica napus. *Mol. Plant*, **14**, 604–619.

Zhang, X., Shiu, S., Cal, A. and Borevitz, J.O. (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genet.*, **4**.

Zhang, X., Yazaki, J., Sundaresan, A., et al. (2006) Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*, **126**, 1189–1201.

Zhao, C., Zayed, O., Zeng, F., et al. (2019) Arabinose biosynthesis is critical for salt stress tolerance in Arabidopsis. *New Phytol.*, **224**, 274–290.

Zhao, M., Zhang, B., Lisch, D. and Ma, J. (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell*, **29**, 2974–2994.

Zhou, X., Liao, Y., Kim, S.U., Chen, Z., Nie, G., Cheng, S., Ye, J. and Xu, F. (2020) Genome-wide identification and characterization of bHLH family genes from Ginkgo biloba. *Sci. Rep.*, **10**, 1–15.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.*, **151**, 3–15.

Reconstruct 100 million years of grass genome evolution through polyploidization

Summary

Polyploidy, the result of whole genome duplication, is a ubiquitous phenomenon in plants and a major force for adapting to their environment. Understanding its mechanisms has benefited from the advent of plant sequencing. However, theories on the rules that govern the evolution of polyploid genomes differ among authors. This study proposes an analysis of the mechanisms driven by polyploidy in a panel of 8 Poaceae species over the last 100 million years. The comparative analysis of the omics data available for these species reveals the impacts on the structure, the fate of duplicated genes and the regulation of genomes. These results allow us to propose a post-polyploidy evolutionary scenario for cereals, identifying the major events and their temporal sequence. They demonstrate that post-polyploidy genomic reprogramming is more complex than what is traditionally reported in the study of a single species.

Reconstruire 100 millions d'années d'évolution des génomes de graminées par polyploïdisation

Arnaud BELLEC

Thèse présentée et soutenue à Castanet-Tolosan, le 15 décembre 2022

Unité de recherche : CNRGV - Centre National de Ressources Génomiques Végétales

Spécialité : Développement des plantes, interactions biotiques et abiotiques

Thèse dirigée par Nicolas LANGLADE et Jérôme SALSE

Résumé

La polyploïdie, résultat de la duplication du génome entier, est un phénomène omniprésent chez les plantes et force majeure pour l'adaptation à leur environnement. La compréhension de ses mécanismes a bénéficié de l'avènement du séquençage des plantes. Cependant, les théories sur les règles qui régissent l'évolution des génomes polyploïdes divergent selon les auteurs. Cette étude propose une analyse des mécanismes actionnés par la polyploïdie sur un panel de 8 espèces de *Poaceae,* depuis 100 millions d'années. L'analyse comparative des données omiques disponibles pour ces espèces révèle les impacts sur la structure, le destin des gènes dupliqués et la régulation des génomes. Ces résultats permettent de proposer un scénario évolutif post-polyploïdie pour les céréales, identifiant les évènements majeurs et leur séquence temporelle. Ils démontrent clairement que la reprogrammation génomique post-polyploïdie est plus complexe que ce qui est traditionnellement rapporté dans l'étude d'une espèce.

Mots clés

Mots-clés : plante, polyploïdie, évolution, génomique, séquençage