La recherche translationnelle chez le blé tendre : comprendre l'évolution de son génome pour améliorer ses caractères agronomiques

Caroline Pont

UNIVERSITE BLAISE PASCAL **UNIVERSITE D'AUVERGNE**
N°D.U.**2732** ANNEE 2016

ECOLE DOCTORALE SCIENCES DE LA VIE, SANTE, AGRONOMIE, ENVIRONNEMENT
N° d'ordre 697

T h e s e :

Présentée à l'Université Blaise Pascal
pour l'obtention du grade de

D O C T E U R D ' U N I V E R S I T E
(Spécialité : Physiologie et Génétique Moléculaire)

soutenue le 06 octobre 2016

Caroline Pont

LA RECHERCHE TRANSLATIONNELLE CHEZ LE BLE TENDRE : COMPRENDRE L'EVOLUTION DE SON GENOME POUR AMELIORER SES CARACTERES AGRONOMIQUES

Président :	Thierry LANGIN	Directeur de recherche HDR
Membres :	Helene BERGES	Ingénieur de recherche Phd
	Bertrand DUBREUCQ	Directeur de recherche HDR
	Jérôme SALSE	Directeur de recherche HDR Directeur de thèse
Rapporteurs :	Dominique THIS	Professeur HDR
	Richard SIBOUT	Chargé de Recherche HDR

UMR 1095 INRA-UBP Génétique, Diversité et Ecophysiologie des Céréales

REMERCIEMENTS

Je tiens tout d'abord à remercier les membres du Jury d'avoir accepté d'évaluer cette thèse, particulièrement les rapporteurs Dominique This et Richard Sibout, les membres du comité de thèse et Thierry Langin, Directeur de l'Unité INRA GDEC.

Je tiens à remercier Jérôme Salse, Directeur de thèse, responsable de l'équipe Paléo-EVO et mon encadrant depuis plus de 10 ans… Plus qu'un responsable, il m'a montré la direction à suivre, et comment suivre le chemin; j'ai beaucoup appris à ses côtés et j'en apprendrais encore… Je le remercie de sa confiance et de ces discussions, toujours vives et argumentées, mais où les hypothèses ont leur place. La recherche nécessite des idées (et un peu de créativité) ; merci de les laisser s'exprimer. Je me rappelle les discussions lors des premières analyses de duplication des génomes en 2006 ; les tableaux noirs de 'gribouillis' pour la reconstruction de l'ancêtre… 2016 ; ce même tableau noirci de chiffres pour la réconciliation du modèle du blé tendre… que sera 2026 ?

J'en profite pour remercier tous ceux qui ont participé à ces échanges, ce travail d'équipe souvent intense et riche, est très motivant. Je remercie Cécile Huneau, qui a rejoint le binôme paléo-EVO il y a peu de temps, de son aide précieuse, de tout le travail qu'elle fait, et de sa joie de vivre au quotidien. Je remercie Florent Murat, un as de la bio-informatique mais qui sait être pédagogue et disponible. Nos chemins se ressemblent quelque peu (sport & thèse !) et nous avons travaillé ensemble de nombreuses années. Je remercie Moaine El Baidouri de sa contribution (qui n'est pas petite…), et de ses nombreuses discussions, c'est toujours un plaisir d'avoir son avis.

De nombreuses personnes ont participé aux travaux de l'équipe, je ne les citerai pas toutes, mais je ne les oublie pas (elles se trouveront plus bas…). Je n'oublie pas non plus ma famille, mon mari, mes amis, de m'avoir soutenu alors que je n'avais pas choisi le plus court chemin…

ABRÉVIATIONS

aDNA : ancient deoxyribonucleic acid
AGK : ancestral grass karyotype
ATK : ancestral triticeae karyotype
BBSRC : biotechnology and biological sciences research council
BAC : bacterial artificial chromosome
BGI : beijing genomics institute
BLAST : basic local alignment search tool
cM : centimorgan
COS : conserved orthologous set
DD : degree day
ELD : expression level dominance
EST : expressed sequence tag
FAO : food and agricultural organization
FDR : false discovery rate
INRAP : institut national de recherches archéologiques préventives
IRGSP : international rice genome sequencing project
IWGSC : international wheat genome sequencing consortium
kb : kilobase
Mb : mégabase
MITEs : miniature inverted repeat transposable elements
MYA : million years ago
QTL : quantitative trait loci
RNA-seq : RNA sequencing
RPKM : reads per kilobase per million
siRNA : small interfering RNA
SNP : single nucleotide polymorphism
SSCP : single strand conformation polymorphism
SSR : simple sequence repeat
TE : transposable element
UTR : untranslated region
WGD : whole genome duplication
WGS : whole genome shotgun

TABLE DES MATIERES

CONTEXTE DE LA THESE

J'exerce mon activité de recherche au sein de l'équipe PaléoEVO (Paléogénomique & Evolution) depuis sa création en 2006. Au sein de cette équipe, j'ai en charge les projets de recherche de 'génomique translationnelle du blé', comme il sera détaillé dans ce manuscrit. Après un DUT 'génie biologique', j'ai été reçue au concours de l'INRA fin 2003. Ce choix m'a permis de continuer à me former à la recherche, par la recherche, validé par la participation à 22 articles scientifiques mentionnés ci-dessous. Mon parcours professionnel au sein de l'INRA m'a permis de valider les compétences acquises par une licence de génomique (ENCPB, PARIS) en 2008, puis d'un Master 'Génomique Ecophysiologie et Production Végétale' en 2009. C'est dans ce contexte que j'ai entrepris depuis septembre 2014, une démarche d'obtention de Doctorat afin d'obtenir la reconnaissance académique de mon activité de recherche.

1. Elbadouri M, Murat F, Veyssiere M, Molinier M, **Pont** C*, Salse J*. Reconciliating the evolutionary origin of bread wheat (Triticum aestivum). New Phytologist. Accepted may 2016. ***Co-corresponding authors***.
2. O. B. Dobrovolskaya O, **Pont** C, Orlov Y, Salse J. Development of new SSR markers for homoeologous WFZP gene loci based on the study of the structure and location of microsatellites in gene-rich regions of chromosomes 2AS, 2BS, and 2DS in bread wheat. Russian Journal of Genetics: Applied Research. May 2016, Volume 6, Issue 3, pp 330-337.
3. Burstin J, Salloignon P, Chabert-Martinello M, Magnin-Robert JB, Siol M, Jacquin F, Chauveau A, **Pont** C, Aubert G, Delaitre C, Truntzer C, Duc G. Genetic diversity and trait genomic prediction in a pea diversity panel. BMC Genomics. 2015 Feb 21;16:105.
4. Dobrovolskaya O, **Pont** C, Sibout R, Martinek P, Badaeva E, Murat F, Chosson A, Watanabe N, Prat E, Gautier N, Gautier V, Poncet C, Orlov YL, Krasnikov AA, Bergès H, Salina E, Laikova L, Salse J. FRIZZY PANICLE drives supernumerary spikelets in bread wheat. Plant Physiol. 2015 Jan;167(1):189-99.
5. Florent Murat, Caroline **Pont**, Jérôme Salse, Paleogenomics in Triticeae for translational research, Current Plant Biology, Volume 1, August 2014.
6. Zhang R, Murat F, **Pont** C, Langin T, Salse J. Paleo-evolutionary plasticity of plant disease resistance genes. BMC Genomics. 2014 Mar 12;15:187.
7. Murat F, Zhang R, Guizard S, Flores R, Armero A, **Pont** C, Steinbach D, Quesneville H, Cooke R, Salse J. Shared Subgenome Dominance Following Polyploidization Explains Grass Genome Evolutionary Plasticity from a Seven Protochromosome Ancestor with 16K Protogenes. Genome Biol Evol. 2014 Jan;6(1):12-33.
8. **Pont** C, Murat F, Guizard S, Flores R, Foucrier S, Bidet Y, Quraishi UM, Alaux M, Doležel J, Fahima T, Budak H, Keller B, Salvi S, Maccaferri M, Faure S, Feuillet C, Steinbach D, Quesneville H, Salse J. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Soumis Plant J 2013.
9. Suliman M, Chateigner-Boutin AL, Francin-Allami M, Partier A, Bouchet B, Salse J, **Pont** C, Marion J, Rogniaux H, Tessier D, Guillon F, Larré C. Identification of glycosyltransferases involved in cell wall Synthesis of wheat endosperm. J Proteomics. 2013 Jan
10. Abrouk M, Zhang R, Murat F, Li A, **Pont** C, Mao L, Salse J. Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. Plant Cell. 2012 May;24(5):1776-92.
11. Dibari B, Murat F, Chosson A, Gautier V, Poncet C, Lecomte P, Mercier I,Bergès H, **Pont** C, Blanco A, Salse J. Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. BMC Genomics. 2012 Jun 6;13:221.
12. **Pont** C, Murat F, Confolent C, Balzergue S, Salse J. RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (Triticum aestivum L.). Genome Biol. 2011 Dec 2;12(12):R119.

13. Quraishi UM; Murat F; **Pont** C; Foucrier S; Desmaizieres G; Confolent C; Rivière N; Charmetg; Paux E; Murigneux A; Guerreiro L; Lafarge S; LeGouis J; Feuillet C; Salse J. Cross-Genome Map Based Dissection of a Nitrogen Use Efficiency Ortho-metaQTL in Bread Wheat Unravels Concerted Cereal Genome Evolution. Plant J. 2011 Mar;65(5):745-56.
14. Dobrovolskaya O, Boeuf C, Salse J, **Pont** C, Sourdille P, Bernard M, Salina E. Microsatellite mapping of Ae. speltoides and map-based comparative analysis of the S, G, and B genomes of Triticeae species. Theor Appl Genet. 2011 Nov;123(7):1145-57.
15. Murat F, Xu JH, Tannier E, Abrouk M, **Pont** C, Messing J, Salse J. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res. 2010 Nov;20(11):1545-57.
16. Quraishi UM, Murat F, Abrouk M, **Pont** C, Confolent C, Oury FX, Ward J, Boros D, Gebruers K, Delcour JA, Courtin CM, Bedo Z, Saulnier L, Guillon F, Balzergue S, Shewry PR, Feuillet C, Charmet G, Salse J. Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (Triticum aestivum L.). Funct Integr Genomics. 2011 Aug 10.
17. Abrouk M, Murat F, **Pont** C, Messing J, Jackson S, Faraut T, Tannier E, Plomion C, Cooke R, Feuillet C, Salse J. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends Plant Sci. 2010 Sep;15(9):479-87.
18. Quraishi UM, Abrouk M, Bolot S, **Pont** C, Throude M, Guilhot N, Confolent C, Bortolini F, Praud S, Murigneux A, Charmet G, Salse J. Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. Funct Integr Genomics. 2009 Nov;9(4):473-84.
19. Throude M, Bolot S, Bosio M, **Pont** C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, Murigneux A, Charmet G, Perez P, Salse J. Structure and expression analysis of rice paleo duplications. Nucleic Acids Research, 2009. 37: 1248-1259.
20. Salse J, Chagué V, Bolot S, Magdelenat G, Huneau C, **Pont** C, Belcram H, Couloux A, Gardais S, Evrard A, Segurens B, Charles M, Ravel C, Samain S, Charmet G, Boudet N, Chalhoub B. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. BMC Genomics. 2008 Nov 25;9:555
21. Jestin L, Ravel C, Auroy S, Laubin B, Perretant MR, **Pont** C, Charmet G. Inheritance of the number and thickness of cell layers in barley aleurone tissue (Hordeum vulgare L.): an approach using F2-F3 progeny. Theor Appl Genet. 2008 May;116(7):991-1002
22. Ravel C, Nagy IJ, Martre P, Sourdille P, Dardevet M, Balfourier F, **Pont** C, Giancola S, Praud S, Charmet G. Single nucleotide polymorphism, genetic mapping, and expression of genes coding for the DOF wheat prolamin-box binding factor. Funct Integr Genomics. 2006 Oct;6(4):310-21.

Le blé hexaploïde, un modèle pour l'étude de la polyploïdie chez les plantes

INTRODUCTION

1. Contexte socio-économique

Le blé est une céréale d'importance économique majeure puisque d'après la FAO sa production mondiale représente 29 % des céréales produites dans le monde. Au total, plus de 715 millions de tonnes ont été produites en 2014 au niveau mondial, faisant du blé la seconde céréale la plus cultivée, derrière le maïs. Avec une consommation mondiale moyenne de 77 kg par personne en 2014, le blé est la céréale la plus consommée devant le riz. Dans l'alimentation humaine, il joue un rôle capital du fait de sa teneur en protéines (approximativement 12 %) et en France, environ 58 % de la production de blé est destinée à l'alimentation humaine (pain, biscuiterie, farine…). L'amélioration du rendement de blé sans perte de qualité est un défi mondial comme l'explique clairement la revue 'The Economist' dans le cadre de son rapport exceptionnel sur l'avenir de l'alimentation mondiale 'the 9 billion question' (The Economist, Feb. 2011). En effet, la production de blé est un réel enjeu pour la population mondiale. Elle qui va passer de 7 milliards d'habitants en 2011 à 9 milliards d'ici 2050, nécessitant une hausse de la production annuelle d'environ 130 Mt (>20 %), et ceci simplement pour garantir aux populations les standards actuels de consommation alimentaire. La problématique réside dans le fait que les stocks mondiaux de blé et de céréales tendent à diminuer. En effet, toujours selon le rapport 'the 9 billion question', les rendements ont augmenté en moyenne de 3 % par an entre 1961 et 1990, période pendant laquelle la population s'est accrue de 1,8 % par an. Mais, entre 1990 et 2007, bien que l'augmentation de la population mondiale ait légèrement diminué à 1.4 %, le rendement annuel de blé n'est quasi plus en augmentation (0,5 %). Cette production de blé est d'autant plus cruciale que les oscillations sont réelles et les stocks mondiaux, même s'ils sont en hausse ces dernières années (*cf.* Figure 1), demeurent restreints (avec selon la FAO une réserve disponible de 3 mois seulement).

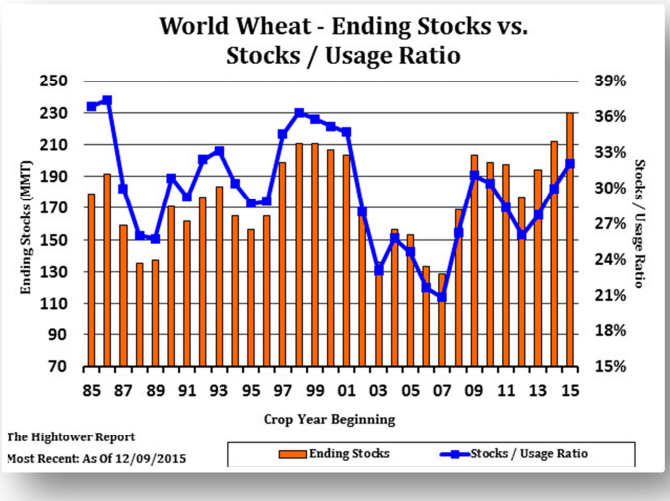


Figure 1. Evolution des stocks mondiaux de blé.
Le graphique illustre l'évolution des stocks mondiaux de blé de 1985 à 2015. La réserve disponible est représentée en orange sur l'échelle de gauche (MMT ; million metric tons), et le ratio (calculé entre le stock et l'utilisation), est représenté en bleu sur l'échelle de droite (en pourcentage). Une chute conséquente est observée entre 2003 et 2008. *Source : www.optionsellers.com*

Certaines années, la demande en blé supérieure à la production mondiale, surtout depuis 2002 (*cf.* Figure 1), maintient la réserve à des niveaux bas (25 %) et cela se traduit par des prix élevés. La

dégradation et le recul des sols arables, associés aux modifications des pratiques alimentaires (notamment la consommation croissante de viande, sachant qu'un kg de poulet nécessite 4 kg de céréales) aboutissent à la gestion en flux tendu des stocks mondiaux de céréales. La réalité est telle, qu'une chute dans la production à l'exemple de la crise de 2007-2008, provoque des hausses de prix records (*cf.* Figure 2) et fait plonger dans la pauvreté des millions de personnes dans les pays en développement qui engagent plus de la moitié de leur revenu à se nourrir. La crise de 2007-2008 a été initiée par des épisodes de sécheresse en Australie et au Canada dans un contexte de forte demande (du fait notamment de l'apparition des agrocarburants), et s'est conclue par des émeutes dans plusieurs pays (*http://www.scienceshumaines.com*). De la même manière, la sécheresse de l'année 2012 a dévasté une grande partie des champs de céréales (notamment aux Etats-Unis) dont le prix a atteint des records (*cf.* Figure 2).

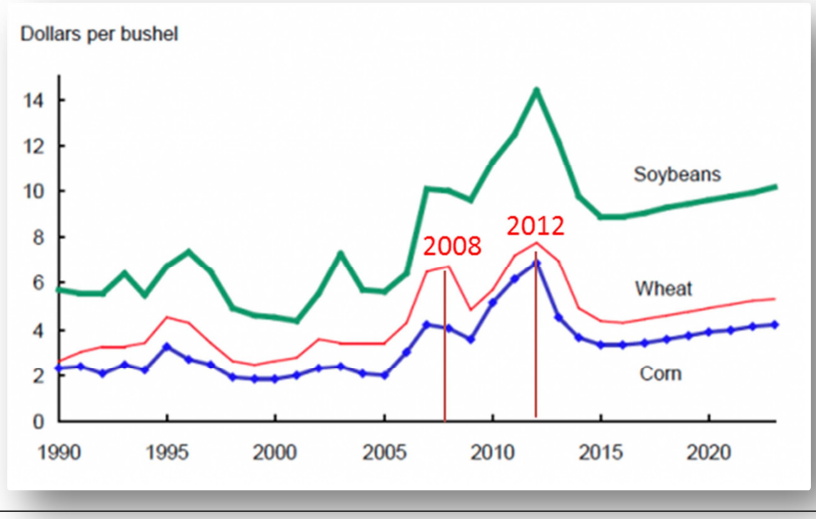


Figure 2. Evolution des prix mondiaux de blé, maïs et soja.
Le graphique illustre l'évolution des prix de 1990 à 2020 de blé (courbe rouge), de maïs (courbe bleue) et de soja (courbe verte), en dollars par boisseau (correspond à 0,03 tonnes). *Source : www.cornandsoybeandigest.com.*

Dans ce contexte, la France a fait de la lutte contre la volatilité des cours des matières premières, y compris les produits agricoles, une des priorités de sa présidence du G20 en 2011. La France, justement, est le premier exportateur européen de blé ; sur les quelques 37,5 Mt récoltées en 2014, environ 18 millions (48%) sont exportées, ce qui représente une ressource considérable pour la France. Les blés français ont un très bon rendement. La récolte 2014 a vu une hausse de production de plus de 3,4 %, soit la deuxième récolte nationale jamais réalisée ces quinze dernières années, mais le rendement 2014 lui, est seulement le 5ème meilleur enregistré dans l'Hexagone sur les 20 dernières années. Ce sont donc la hausse des surfaces récoltées et les conditions météorologiques printanières qui expliquent ces résultats. Cependant, de forts rendements contribuent à un appauvrissement de la qualité de la récolte du fait de la chute de la teneur en protéines du grain associée, et ceci lié à la quantité limitante d'azote absorbée et remobilisée vers le grain. François Gatel, directeur de France Export Céréales (association à but non lucratif constituée et financée par les producteurs français), explique que la concurrence à l'export est rude et surtout que la qualité des blés français n'est pas toujours à la hauteur. Les cahiers des charges fixent des minimas pour les critères d'humidité, de poids spécifique, de taux de protéines

(11,5 %), d'alvéographes (mesure de la ténacité, de l'extensibilité et de l'élasticité d'un pâton de farine) ou de germination sur pied. Le taux de gluten devient également un critère essentiel notamment pour l'Afrique subsaharienne et le Moyen-Orient. Concernant l'utilisation de blé pour la production de bioéthanol, elle est en augmentation depuis 2004, d'un facteur 'quatre', et repose exclusivement sur la teneur en amidon du grain.

Le défi mondial est donc d'améliorer considérablement le rendement du blé tendre (de plus de 20 % sur les 100 prochaines années), d'assurer sa stabilité au regard des changements climatiques, et de maintenir la qualité du grain pour répondre aux besoins des marchés, et ce, en préservant l'environnement. Le rendement annuel moyen est de 74 quintaux par hectare en France, mais varie considérablement selon les conditions environnementales ; il est par exemple de 90 q/ha dans le Pas-de-Calais et de 50 q/ha dans le Gers. Ce rendement est le fruit de gros efforts en sélection, réalisés depuis les années 60. De nouveaux herbicides ont été l'introduits lors de la révolution verte, puis de nouveaux fongicides, associés à des variétés à haut rendement notamment grâce aux travaux de Norman Borlaug, prix Nobel de la paix en 1970. Après deux décennies de travaux en collaboration avec des scientifiques mexicains, il développe une nouvelle variété de blé semi-naine qui a permis de doubler la production de blé en Inde et au Pakistan entre 1965 et 1970. Elle a été par la suite introduite dans toute l'Amérique latine, au Proche-Orient et en Afrique. Concernant la qualité du grain, d'après Arvalis (institut du végétal et référent sur la recherche et l'accompagnement technique du plan national « Protéines Blé»), les leviers connus pour améliorer la teneur en protéines des blés sont notamment la variété (enjeu de 0,5 à 1 %) et l'apport d'azote (de 0,5 à 2 %), alors que les facteurs non maîtrisés (climat, maladies…) peuvent aussi avoir un impact conséquent (+/- 0,5 à 2 %).
Quelles démarches mettre en œuvre pour relever le défi considérable de l'augmentation de la production nécessaire dans les années futures ? La solution la plus immédiate serait d'augmenter la superficie cultivée. En France, il est souvent discuté du risque d'un manque de terres aptes à l'agriculture dans les années à venir, mais d'après les projections de la FAO, ce ne sera pas le cas au niveau mondial. En effet, bien que dans certaines régions les pénuries risquent fort de s'aggraver, la superficie arable dans les pays en développement va augmenter de près de 13 %, soit 120 millions d'hectares au cours de la période 2000 à 2030. Ceci amènera sans doute des changements drastiques des modes de culture et le développement de variétés cultivables sur tous types de sols (c.-à-d. résistants aux contraintes environnementales). Dans ce contexte, la recherche française mise notamment sur les connaissances fondamentales de la biologie du blé et notamment sur l'étude de son génome. L'objectif est de comprendre l'organisation, le fonctionnement, la régulation et l'évolution de ce génome complexe pour décrypter les mécanismes contrôlant les caractères agronomiques importants (rendement, qualité du grain, résistance à des stress biotiques et abiotiques). Les travaux de thèse présentés dans ce manuscrit s'inscrivent dans ce contexte général.

2. L'essor de la génomique chez le blé

La génomique est tournée vers la compréhension du fonctionnement et de l'organisation des gènes. Elle peut être abordée en deux approches : la génomique structurale qui décrit l'organisation des chromosomes, des gènes et leurs polymorphismes ; et la génomique fonctionnelle qui a pour objectif d'attribuer une fonction biologique à ces gènes ainsi que de décrypter la façon dont ils sont régulés. La

Figure 3 illustre l'essor de la génomique, notamment chez le blé, ayant mené au séquençage de son génome.

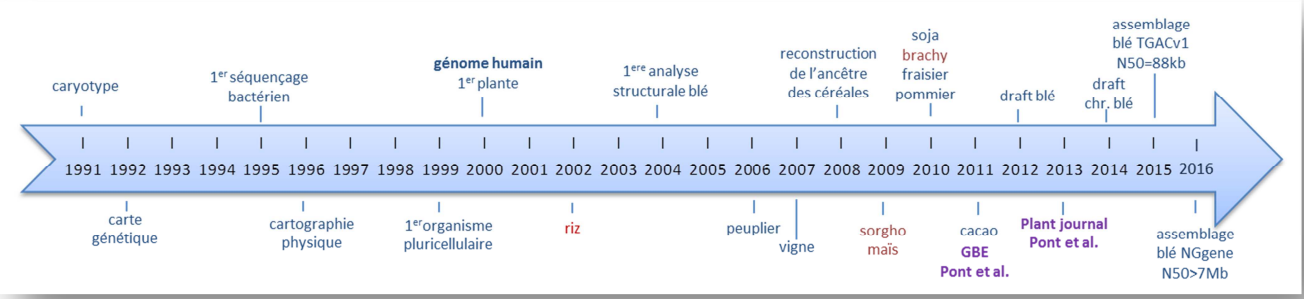


Figure 3. Essor de la génomique du blé.
Les dates marquantes de l'essor de la génomique du blé sont illustrées dans ce diagramme avec notamment l'évolution du séquençage des génomes comme détaillé dans le texte. Les espèces modèles 'monocotylédones' sont matérialisées en rouge et les 2 articles présentés dans ce manuscrit, en violet.

L'ère de la génomique a été initiée par le programme international de séquençage du génome humain lancé dès 1985. Trois milliards d'euros furent investis à partir de 1989 pour l'immense projet de séquençage des 24 chromosomes (3,2 milliards de nucléotides). Cependant, Craig Venter, démontre en 1995 que le séquençage aléatoire pourrait être appliqué à des génomes entiers avec rapidité et précision. *Via* le séquençage complet 'Whole Genome Shotgun' (WGS), il prend l'exemple d'une bactérie (Haemophilus influenzae, 1 830 137 bp, Fleischmann *et al.* 1995). Fort de ce succès, il s'en suit une course au séquençage des génomes (*cf.* Figure 4) et le premier organisme pluricellulaire à avoir été entièrement séquencé est le nématode *Caenorhabditis elegans* avec 97 Mb et 19 099 gènes (Sequencing Consortium *et al.* 1999). Puis très rapidement, viendra la drosophile (Myers *et al.* 2000) avec 13 601 gènes et 180 Mb. Concernant le projet de séquençage du génome humain, un premier brouillon (dit 'draft') a été publié en 2000. Il comportait encore un grand nombre de trous (dit 'gaps') et d'imperfections. La séquence complète n'a été achevée qu'en 2004 par le consortium international public (International Human Genome Sequencing Consortium 2004). La même année que le génome humain, fut séquencée la première plante, *Arabidopsis thaliana,* une dicotylédone (Arabidopsis Genome Initiative 2000). Elle avait été choisie en 1980 comme espèce de référence, notamment du fait de son cycle de reproduction court, sa petite taille et son petit génome organisé en cinq paires de chromosomes comprenant 125 Mpb et 27 228 gènes. Elle restera, pour toute la communauté scientifique, l'espèce modèle de référence chez les plantes. La seconde plante dont le génome sera décrypté est le riz, *Oryza sativa*, où la course au séquençage est montée à son apogée. Deux articles relatifs au séquençage de deux sous-espèces *via* WGS, effectué par le BGI (Beijing Genomics Institute) et Syngenta, sont parus dans le même numéro de Science en avril 2002 (Yu J. *et al.* 2002 et Goff S. *et al.* 2002), alors que le consortium international annoncera avoir obtenu la séquence du riz de haute qualité (couverture 10X en moyenne) fin 2002 (Sasaki T. *et al.* 2002 ; Feng Q. *et al.* 2002). Cette séquence de meilleure qualité (IRGSP 2005) constituera la référence chez les monocotylédones et permettra une meilleure compréhension de la structure du génome du riz avec la localisation précise des gènes sur les pseudomolécules (séquences génomiques entièrement assemblées couvrant chaque chromosome). Depuis le séquençage

d'*Arabidopsis* et du riz, un grand nombre de génomes de plantes a été séquencé, permettant l'étude des différentes familles botaniques des angiospermes (plantes à fleurs). Plus de 100 génomes de plantes ont été séquencés, dont une cinquantaine depuis 2013, couvrant la totalité de la diversité phylogénétique des angiospermes (*cf.* Figure 4).

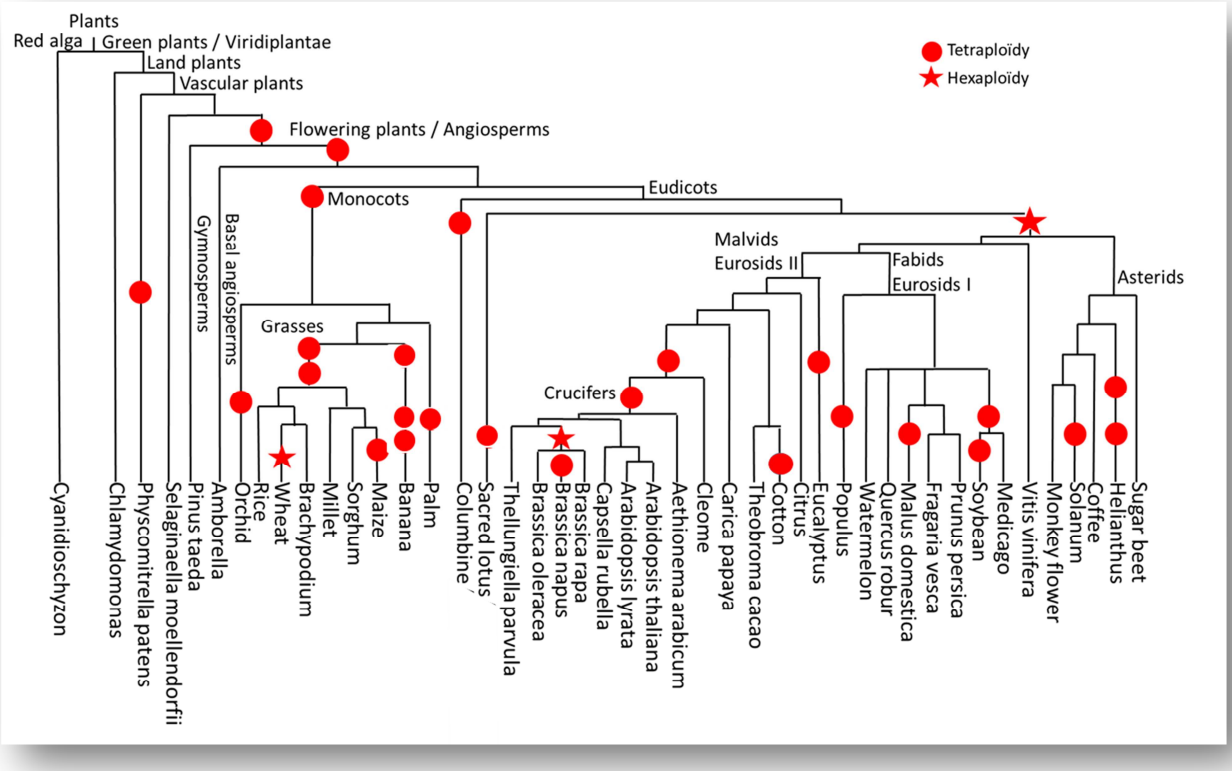
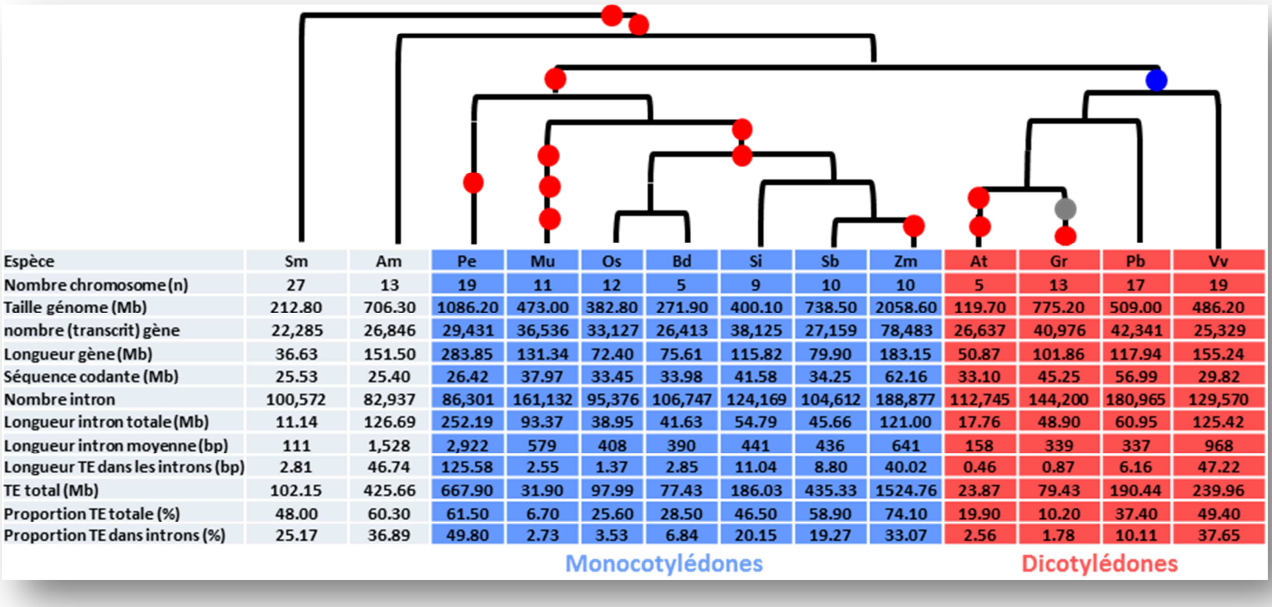


Figure 4. Arbre phylogénétique des angiospermes illustrant les génomes séquencés et les évènements de polyploïdie
Cette figure illustre l'arbre phylogénétique des espèces de plantes à fleurs dont le génome est séquencé en 2013. Les cercles et étoiles rouges représentent respectivement les évènements de tétraploïdie et d'hexaploïdie. *Adaptée de Cogepedia https://genomevolution.org/coge/*

Les génomes d'angiospermes à disposition mettent en évidence une très grande diversité structurale (Cai *et al.* 2014). Les caractéristiques génomiques montrent la diversité des génomes d'angiospermes, telles que leur structure caryotypique, leur taille, ou leur contenu en gènes et en éléments transposables 'TE' (Cai *et al.* 2014). A titre d'exemple, le génome d'*Arabidopsis thaliana* est composé de 5 chromosomes pour 120 Mb et 20% de TE, alors que celui de *Zea Mays* (maïs) comporte 10 chromosomes pour 2 Gb et 74% de TE (*cf.* Figure 5).



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

Figure 5. Diversité structurale des génomes de plantes à fleurs

Cette figure illustre la diversité structurale des génomes d'Angiospermes, à travers l'analyse de leurs caractéristiques génomiques. Le tableau résume les principales caractéristiques génomiques des plantes à fleurs. L'arbre phylogénétique des espèces étudiées dans le tableau est représenté en haut de la figure. Les points rouges représentent les évènements de duplication, le point bleu, l'évènement de triplication et le point gris, de sextuplication. 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. A*dapté de Cai et al. 2014*

Concernant le blé tendre, malgré l'importance économique de cette céréale, le séquençage de son génome a relativement tardé (*cf.* Figure 5) du fait de sa complexité; notamment sa taille, sa richesse en séquences répétées et les duplications de son génome. Avec ses 15 Gb répartis sur 21 chromosomes, le génome du blé est le plus grand de toutes les céréales connues (cinq fois plus grand que le génome humain). Il est principalement constitué de séquences répétées non codantes, les TE, qui représentent plus de 80 % du génome du blé (Charles *et al.* 2008) et qui rendent l'assemblage des séquences difficile. La Figure 6 résume les différents séquençages réalisés à ce jour. Dans l'objectif du séquençage du génome du blé, le consortium IWGSC (International Wheat Genome Sequencing Consortium) a été créé en 2005 avec plus de 1000 partenaires, représentant 57 pays. L'approche choisie par le consortium est basée sur le séquençage BAC à BAC (bacterial artificial chromosome ; fragment génomique d'environ 100 kb inséré dans une bactérie), *via* la construction de banques ordonnées et spécifiques après le tri des 21 chromosomes (par cytométrie de flux). L'essor des nouvelles technologies de séquençage (en termes de débit et de taille de séquences) est tel, qu'à la marge de ce consortium, des scientifiques britanniques du BBSRC (www.bbsrc.ac.uk) ont rendu publiques en 2010 les premières séquences génomiques de blé tendre obtenues par WGS. Ce premier séquençage très parcellaire (couverture 1x) sera enrichi pour atteindre une couverture de 5X, et ainsi, Brenchley *et al.* 2012 publieront le premier séquençage partiel du génome du blé délivrant 94 000 gènes répartis sur les 21 chromosomes. Bien que cet assemblage soit fragmenté, il constitue un support efficace pour la caractérisation de gènes, et le développement de marqueurs moléculaires lors des études génétiques. A la suite de cela en 2014, le consortium publie lui aussi une ébauche WGS, mais par bras chromosomique (99 386 gènes, IWGSC 2014, Bolser *et al.* 2014)

et avec une première pseudomolécule du chromosome 3B de 774Mb (Choulet *et al.* 2014). La course au séquençage continue avec l'augmentation des longueurs de séquences produites, mais aussi, de la qualité de l'assemblage en WGS (*cf.* Figure 6). Ainsi, 'The Genome Analysis Centre' (TGAC, www.earlham.ac.uk) en Angleterre, rend publique en 2015, sa première version du génome assemblé de 13,4 Gb avec une taille médiane de contigs de 89kb (N50). Enfin, en 2016, la société NRgene révolutionne l'essor de la génomique du blé tendre en levant un des principaux verrous techniques ; l'assemblage des séquences répétées. Elle tire profit de l'algorithme DeNovoMAGIC et délivre les 21 pseudomolécules, avec un N50 >7Mb (*cf.* Figure 5), en seulement 3 semaines. Ces deux dernières références sont accessibles à la communauté scientifique depuis Juin 2016 avec une publication prévue pour 2017 (http://nrgene.com/press-releases). Au-delà des différentes séquences obtenues pour le blé tendre hexaploïde, des références parcellaires ont été également obtenues pour les blés diploïdes et tétraploïdes (*cf.* Figure 6).

publication / année	variété	provenance	Nom de l'assemblage	N50 *(kb)*	nb scaffold	nb de gènes
Brenchley et al. 2012	Ta. CS (ABD)	BBSRC	CS5X	1	-	20 496
Ling et al. 2013	Triticum urartu (A)	BGI	T.ura-BGI	64	19 000	34 879
Jia et al. 2013	Aegilops tauschi (D)	BGI	Ae.ta-BGI	58	19 000	50 264
IWGSC 2014	Ta. CS (ABD)	IWGSC	CSS	2	>1 000 000	99 386
Choulet et al. 2014	Ta. CS (ch.3B)	GDEC/CNS	3B-pseudo	892	296	5 326
Chapman et al. 2015	Ta. ITMI pop (ABD)	IPK/JGI	Syn-JGI	21	120 000	NA
Juillet 2015	Wild emmer wheat (AB)	NRGene	WEW-NRGene	7 000	414	NA
Janvier 2016	Ta. CS (ABD)	TGAC/BBSRC	TGACv1	89	43 000	99 386
Juin 2016	Ta. CS (ABD)	NRGene/IWGSC	IWGSC-WGA	7 394	547	NA

Figure 6. Séquences génomiques de blé disponibles à ce jour, par ordre chronologique de séquençage.
Ce tableau liste les données génomiques disponibles par ordre chronologique chez le blé diploïde (en bleu), tétraploïde (en rose) et hexaploïde (en vert), avec leur provenance, la taille médiane des contigs (N50 en kb) et le nombre de scaffold contenus. *Abréviations : Ta : triticum aestivum ; CS : Chinese Spring ; WGA : Whole genome assembly ; ITMI : International Triticeae Mapping Initiative ; BBSRC : Biotechnology and Biological Sciences Research Council ; BGI : Beijing Genomics Institute ; IWGSC : International Wheat Genome Sequencing Consortium ; CNS : Centre National de Séquençage ; GDEC : Génétique, Diversité et Ecophysiologie des Céréales ; IPK : Leibniz-Institute of Plant Genetics and Crop Plant Research ; JGI : Joint Genome Institute ; TGAC : The Genome Analysis Centre.*

3. Histoire évolutive des céréales et du blé tendre hexaploïde

La trace la plus ancienne de céréales date de 56 à 71 millions d'années en Inde, sous forme de phytolithes, particules cristallines provenant de l'accumulation de silice à l'intérieur des plantes (Piperno *et al.* 2005). Dans ces mêmes travaux, l'examen microscopique d'excréments fossilisés de dinosaures a montré qu'ils se nourrissaient d'au moins cinq types différents de plantes. En réalité, il y a environ 150 millions d'années, les plantes à fleurs ont subi une différenciation en deux grands groupes : les dicotylédones et les monocotylédones. Parmi ces dernières, on retrouve la famille des *Poaceae* (riz sorgho, maïs…), apparue il y a environ 60 millions d'années. Le blé tendre, *Triticum aestivum*, est né au sein de cette famille à l'état sauvage il y a 10 000 ans au Moyen-Orient. Il a cependant fallu attendre 8 000 avant JC pour qu'il soit domestiqué et cultivé par l'homme.

Paléogénomique des céréales

La paléogénomique vise à étudier la structure et la fonction du génome ancestral à l'origine des espèces modernes, et ainsi comprendre leur histoire évolutive. Deux approches permettent d'aborder cette thématique ; la première par le séquençage d'ADN fossile quand celui-ci est disponible. Ce fut le cas pour le mammouth (Poinar *et al.* 2006) et l'homme de Neandertal (Green *et al.* 2010). La seconde approche consiste à comparer les génomes d'espèces actuelles. Dans ce dernier cas, l'identification des gènes conservés entre espèces modernes permet de reconstruire le génome ancestral (génome minimal théorique dans sa structure chromosomique et son contenu en gènes). La reconstruction des génomes ancestraux à partir de la comparaison des génomes modernes permet de proposer un scenario évolutif le plus parcimonieux, c'est-à-dire en introduisant le plus petit nombre de réarrangements (inversions, délétions, translocations, duplications) nécessaires pour expliquer la transition entre les génomes ancêtres et modernes.

La première grande étude structurale du génome du blé a été menée en 2004 sur 16 000 loci issus d'EST cartographiées (Qi *et al.*2004), dévoilant la répartition des gènes en exploitant des lignées dites de délétions, créées par Sears dès 1954. Ces lignées sont des outils de référence où tout, ou une partie, de chromosome est absent (Endo *et al.* 1996). Ces lignées permettent ainsi de localiser physiquement des marqueurs moléculaires au sein de bins (régions chromosomiques couvrant les délétions). Ces données d'assignation des gènes ont par la suite pu être exploitées pour comparer le génome du blé avec les céréales apparentées dont le génome est séquencé telle que le riz. Ainsi, notre équipe a identifié les relations de synténie (conservation des gènes) et de paralogie (duplication des gènes), ayant permis de proposer le premier modèle évolutif des génomes de céréales à partir d'un ancêtre commun (AGK, *cf.* Figure 7) constitué de 7 chromosomes et porteur de près de 16 464 protogènes (Salse *et al.* 2008a, Bolot *et al.* 2009, Murat *et al.* 2010, 2014). Dans ce scénario, le génome ancestral a subi une duplication totale de son génome (1R en rouge, Figure 7) il y a 90 millions d'années pour atteindre une structure à 14 chromosomes, puis à 12, suite à deux évènements de fusions chromosomiques. Les céréales modernes dérivent toutes de cet ancêtre AGK à 12 chromosomes datant de plus de 65 millions d'années. Le riz moderne a conservé cette structure à 12 chromosomes. Les *Panicoideae,* avec le maïs et le sorgho, ont évolué dans un premier temps (il y a plus de 29 MYA) à partir de l'ancêtre à 12 chromosomes par 2 fusions chromosomiques pour donner un nouvel ancêtre intermédiaire à 10 chromosomes. Le sorgho a gardé cette structure à 10 chromosomes, tandis que le maïs a connu une nouvelle duplication totale de son génome, il y a environ 5 millions d'années, en donnant un ancêtre intermédiaire à 20 chromosomes. Puis, il a connu rapidement 17 nouvelles fusions chromosomiques pour aboutir à la structure moderne de son génome constitué de 10 chromosomes. Quant à *Brachypodium*, ces études suggèrent qu'à partir de l'ancêtre à 12 chromosomes, 7 événements de fusions chromosomiques sont à l'origine de sa structure génomique actuelle à 5 chromosomes. Concernant les *Triticeaes*, il y a moins de 26 MYA, leur ancêtre a subi 5 fusions de chromosomes pour aboutir à l'ancêtre des *Triticeaes* ATK (ancestral triticeae karyotype, *cf. Figure 7*) composé de 7 chromosomes. Depuis, l'orge a maintenu cette structure, mais le blé et le seigle ont subi des événements supplémentaires tels que des translocations et/ou polyploïdisation. Le seigle partage une translocation commune avec le blé et 5 translocations qui lui sont spécifiques. Le blé lui, a connu par la suite 2 cycles de polyploïdisation (2R et 3R, en rouge Figure 8) puis une translocation aboutissant au génome hexaploïde que l'on connaît à l'heure actuelle composé de 21 chromosomes (7x3 chromosomes).

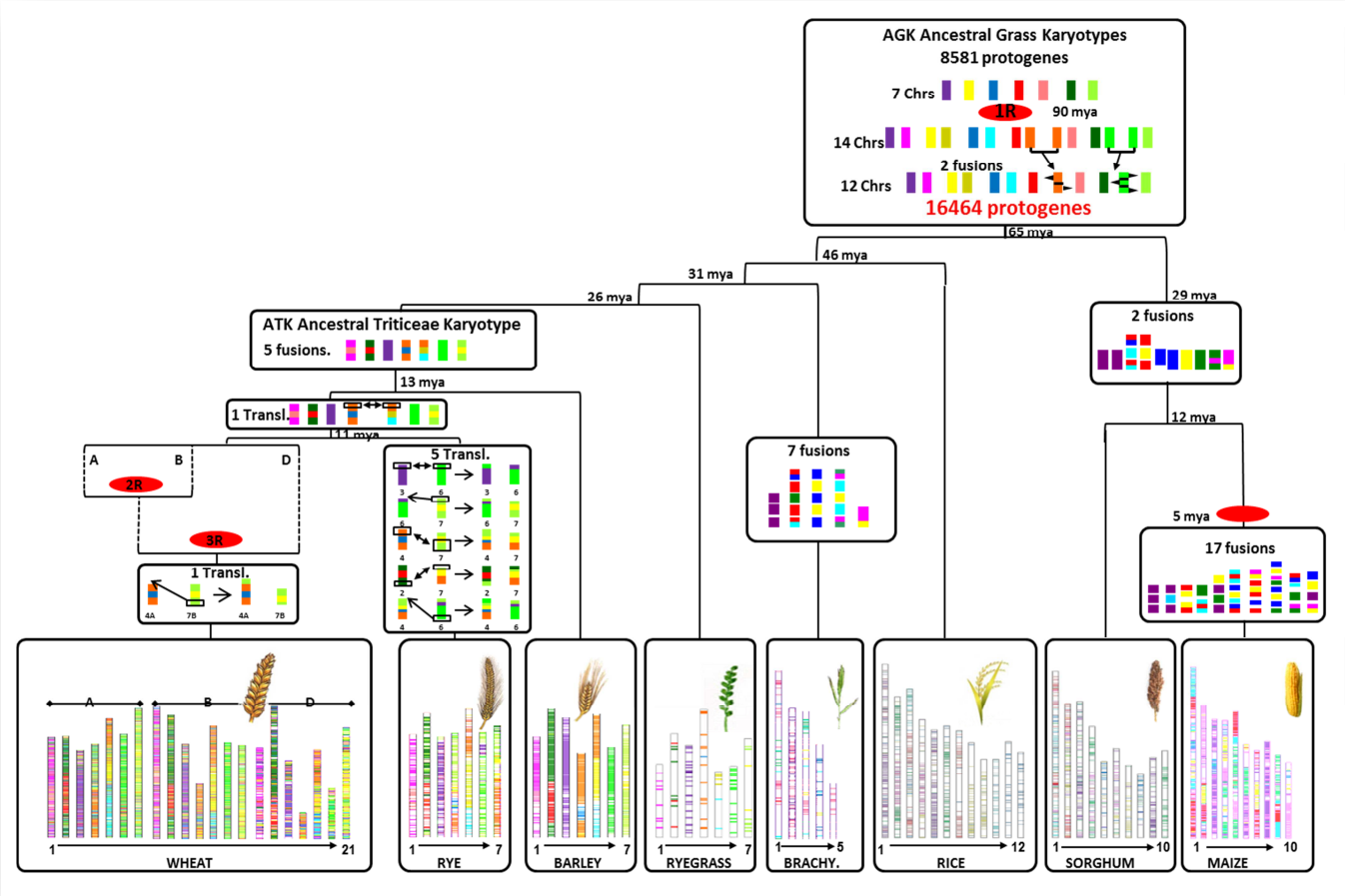


Figure 7. Histoire évolutive des céréales à partir d'un ancêtre commun vieux de 90 millions d'années
Modélisation de l'évolution des génomes des céréales à partir d'un ancêtre commun AGK (Ancestral Grass Karyotype) à 7 chromosomes. Les chromosomes sont représentés par un code de 7 couleurs provenant de l'origine de l'ancêtre à 7 chromosomes et permettant de retracer leur évolution au sein des génomes modernes (représenté en bas). L'ancêtre des *Triticeae* ATK (Ancestral Triticeae Karyotype) est représenté suite à 5 fusions chromosomiques communes à toutes les *Triticeae*. Le nombre de réarrangements (fusion ou translocation) est indiqué sur chaque branche de l'arbre phylogénétique. Les dates de spéciation des différentes espèces (et ancêtres) sont indiquées en millions d'années (mya). *Adaptée de Murat et al. 2014*

Origine de l'hexaploïdie du blé tendre

Le blé tendre est une espèce hexaploïde résultant de la fusion des génomes de trois progéniteurs diploïdes ; son histoire évolutive a connu au total depuis l'ancêtre des céréales trois cycles de polyploïdisation (*cf.* Figure 8). Le premier événement de polyploïdisation correspond à la paléo-tétraploïdie commune à l'ensemble des céréales comme détaillée dans la section précédente (rond bleu, Figure 8), suivie de deux évènements récents de néo-polyploïdisation (il y a 0,5 et 0,01 MYA, ronds vert et violet, Figure 8). Le blé tendre hexaploïde est issu des progéniteurs A, *Triticum urartu* et D, *Aegilops tauschii*. L'origine du parent donneur du sous-génome B n'a pas été clairement identifiée, mais il apparait être très proche de *Aegilops speltoides* (Salse *et al.* 2008b, Kilian *et al.* 2007). Le blé tendre, principalement consommé sous forme de farine pour le pain est hexaploïde, il dérive du blé dur tétraploïde, consommé lui sous forme de pâtes.

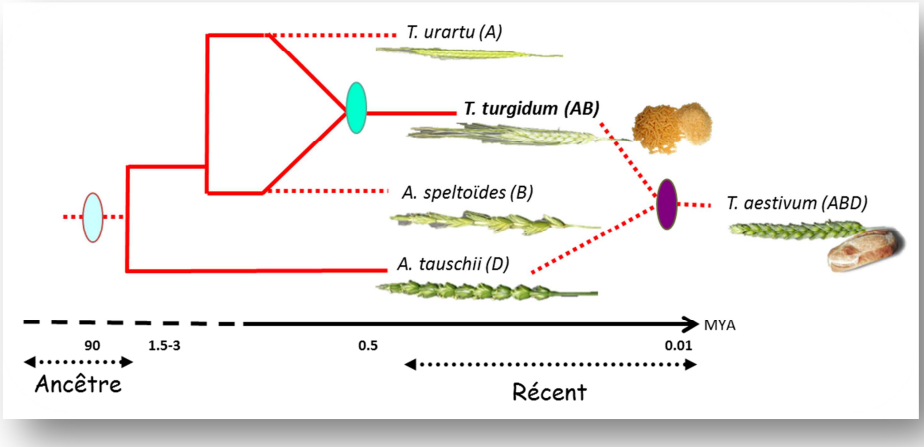


Figure 8. Origine de l'hexaploïdie du blé tendre
Ce schéma illustre l'histoire évolutive du blé tendre au cours du temps. Les 3 trois cycles de polyploïdisation sont matérialisés par les ronds ; bleu il y a 90 MYA, vert il y a 500 000 ans et violet il y a 10 000 ans. Après la duplication ancestrale commune aux céréales, la fusion de trois génomes (ABD) a donné le blé tendre (*Triticum aestivum*) cultivé et consommé sous forme de pains ou de farines. Le génome A est issu de *Triticum urartu*, le B *d'Aegilops speltoides* et le D *d'Aegilops tauschii*. L'intermédiaire (*Triticum turgidum*) correspond au blé dur tétraploïde, espèce cultivée pour la fabrication des pâtes.

Ainsi, sur la base des connaissances acquises de l'évolution des génomes de céréales, pour une région du génome de l'ancêtre des céréales (AGK), six régions sont héritées des trois évènements de polyploïdisation. Ces 6 régions sont donc potentiellement présentes au sein des 21 chromosomes du génome du blé moderne (*cf.* Figure 89, rectangles roses). Les gènes conservés entre le blé et les autres céréales sont appelés orthologues. Les copies de gènes issus de la paléo-tétraploïdisation retenus en 6 copies chez le blé tendre sont appelés ohnologues (*cf.* Figure 89, en violet). Ceux hérités de la néo-hexaploïdisation sont appelés homéologues (*cf.* Figure 89, en bleu).

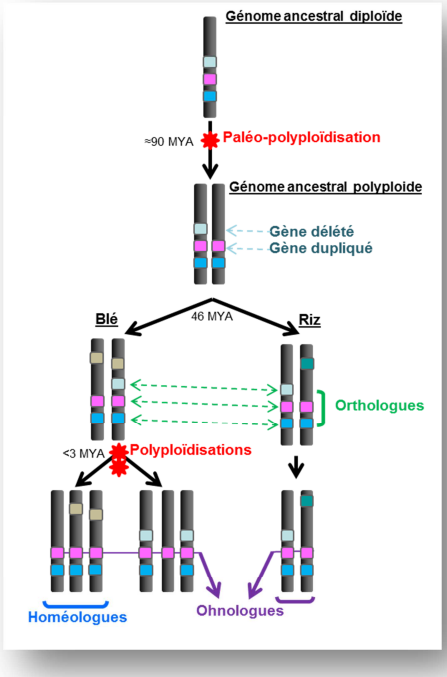


Figure 9. Illustration de relations d'homologie des gènes chez le blé.
La figure illustre l'évolution d'une région ancestrale (haut) en six régions homéologues chez le blé tendre (bas). Le génome ancestral a connu une première duplication il y a 90 millions d'années (MYA), puis, après la spéciation des deux espèces blé et riz, le blé a connu l'hexaploïdie générant jusqu'à six copies de gènes ancestraux (boites de couleur). Lorsque l'on considère deux espèces apparentées, les gènes dérivant d'un même ancêtre sont orthologues et ceux dérivant d'une duplication sont des ohnologues si ils sont conservés en réponse à un évènement de paléo-polyploïdie ancien. A l'intérieur d'une même espèce, les gènes issus d'une polyploïdisation sont dits homéologues.

4. La polyploïdie et ses effets immédiats

La polyploïdie est un processus de doublement du matériel génétique tout à fait naturel. En effet cette duplication globale du génome (WGD) est très fréquente dans le règne végétal (*cf.* Figure 4 ; Doyle et al. 2008; Freeling, 2009) ; 70 à 80 % des plantes à fleurs ont eu un ancêtre polyploïde (paléo-polyploïdie) et 47 % sont encore actuellement polyploïdes comme le blé (Soltis *et al.* 2014). A la suite d'un évènement de polyploïdie, la cellule va acquérir un ou plusieurs jeux de son contenu génomique initial (*cf.* Figure 10). On distingue deux types de polyploïdie ; l'autopolyploïdie (Müntzing, 1936) qui duplique les chromosomes d'une même espèce au sein du noyau par doublement chromosomique et, l'allopolyploïdie (Moore, 2002 ; Van de Peer et al. 2009) menant à un état polyploïde suite à une hybridation entre génomes d'espèces apparentées. Le doublement chromosomique se fait par doublement somatique, ou fusion de gamètes non réduits. S'il n'y a pas doublement chromosomique l'hybride diploïde obtenu est alors homoploïde (*cf.* Figure 10).

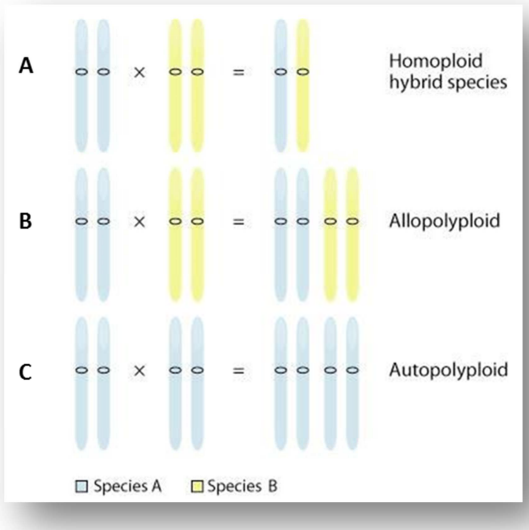


Figure 10. Illustration schématique d'une autopolyploïdisation, d'une allopolyploïdisation et d'une hybridation homoploïde.
L'origine des espèces est matérialisée par une seule paire de chromosomes provenant de chacune des espèces parentales diploïdes (bleu ou jaune). (A) L'hybridation de deux espèces diploïdes donne naissance à un homoploïde portant de façon complémentaire un exemplaire de chaque chromosome parental (diploïde). (B) La formation d'un allopolyploïde est issue elle aussi d'une hybridation entre A et B, cependant, contrairement à l'homoploïde, l'allopolyploïdisation implique le doublement des chromosomes combinant l'ensemble des génomes nucléaires des deux espèces parentales. (C) L'autopolyploïdisation implique également le doublement des chromosomes et se produit par l'intermédiaire d'un croisement des individus diploïdes de la même espèce; le doublement des chromosomes donne alors un autopolyploïde. *Source : Soltis et al. 2009.*

L'autopolyploïdisation est issue de la duplication des chromosomes au sein de la même espèce, c'est le cas pour la pomme de terre (4x - 48 chromosomes), la banane (3x – 33 chromosomes), la cacahuète (4x – 40 chromosomes). Le processus de polyploïdisation le plus répandu demeure l'allopolyploïdisation, c'est le cas pour le coton (4x – 52 chromosomes), le colza (4x – 38 chromosomes), la canne à sucre (8x – 80 chromosomes) et le blé (6x – 21 chromosomes). Dans ce cas, l'hybridation interspécifique produit une nouvelle espèce résultant de deux génomes différents (ou plus), enveloppés dans un même noyau suivi

d'un doublement des chromosomes pour pallier à la stérilité de l'hybride post-hybridation. Il existe des blés diploïdes (2n = 2x = 14 chromosomes), cependant les blés les plus cultivés et consommés dans le monde sont le blé dur tétraploïde (2n = 4x = 28 chromosomes), et surtout le blé tendre hexaploïde (2n = 6x = 42 chromosomes). Ce dernier est composé de 3 jeux de 7 chromosomes. Ces 3 sous-génomes (appelés A, B et D) proviennent de deux évènements de polyploïdisation. La première hybridation a eu lieu il y a environ 500 000 ans, entre le blé diploïde *Triticum urartu* (AA) et un autre diploïde, proche d'*Aegilops speltoides* (BB) ayant donné le blé dur (*Triticum turgidum*, AABB). La deuxième hybridation a eu lieu il y a environ 10 000 ans, entre le blé dur *Triticum turgidum* et le diploïde *Aegilops tauschii* (DD) donnant le blé tendre hexaploïde (*Triticum aestivum*, AABBDD) (*cf.* Figure 11).

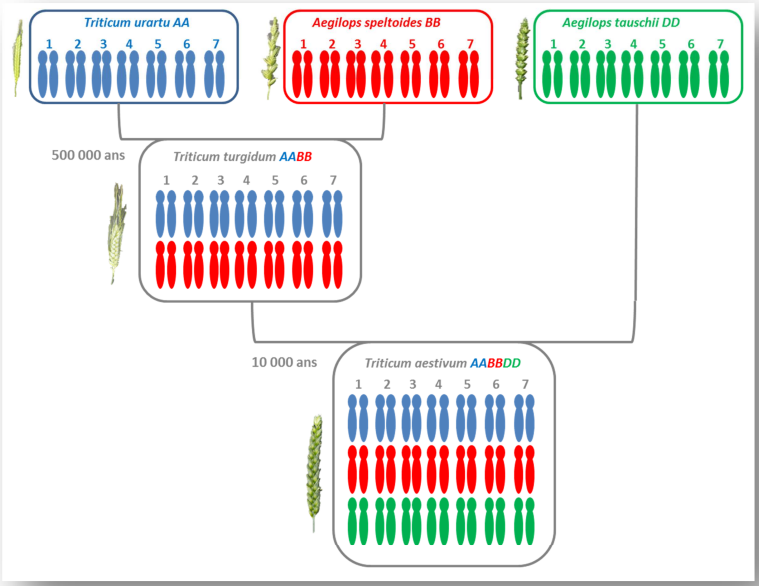


Figure 11. Illustration de l'histoire évolutive du blé tendre hexaploïde

Cette figure représente l'histoire évolutive du blé tendre hexaploïde à travers deux évènements d'allopolyploïdie. Le premier, il y a environ 500 000 ans, entre l'ancêtre *Triticum urartu* (génome A) et l'ancêtre *Aegilops speltoides* (génome B) pour former le blé dur tétraploïde (AB); et le deuxième, il y a environ 10 000 ans, entre l'ancêtre du blé dur (AB) et l'ancêtre *Aegilops tauschii* (génome D) pour former le blé tendre hexaploïde (ABD). Les caryotypes bleus, rouges et verts correspondent respectivement aux caryotypes des ancêtres *Triticum urartu*, *Aegilops speltoides* et *Aegilops tauschii*.

Le blé tendre, est donc allohexaploïde depuis 10 000 ans environ, après deux cycles de WGD récentes. La polyploïdisation est décrite dans la littérature comme une force majeure de l'évolution, car elle aboutit à une spéciation instantanée (Levy *et al.* 2002). Un événement unique de polyploïdisation est suffisant pour établir une barrière taxonomique empêchant tout flux de gènes entre la nouvelle espèce polyploïde, génétiquement isolée, et ses géniteurs. Cet événement conduit donc très souvent à l'isolement reproductif (Soltis *et al.* 2009 ; Madlung *et al.* 2013). Il est établi que la polyploïdie est un processus récurrent de l'évolution des plantes pouvant ainsi permettre aux espèces, post-polyploïdie, de disposer d'une capacité d'adaptation accrue grâce à l'augmentation de la diversité phénotypique issue du choc génomique généré par la polyploïdie (McClintock 1984, Doyle *et al.* 2008, Schranz *et al.* 2004).

Un effet 'gigantisme' des plantes a été observé au niveau des fleurs, des feuilles et des tiges (*cf.* figure 12) chez différentes espèces polyploïdes depuis les années 50 (Elliott 1958, Müntzig, 1961).

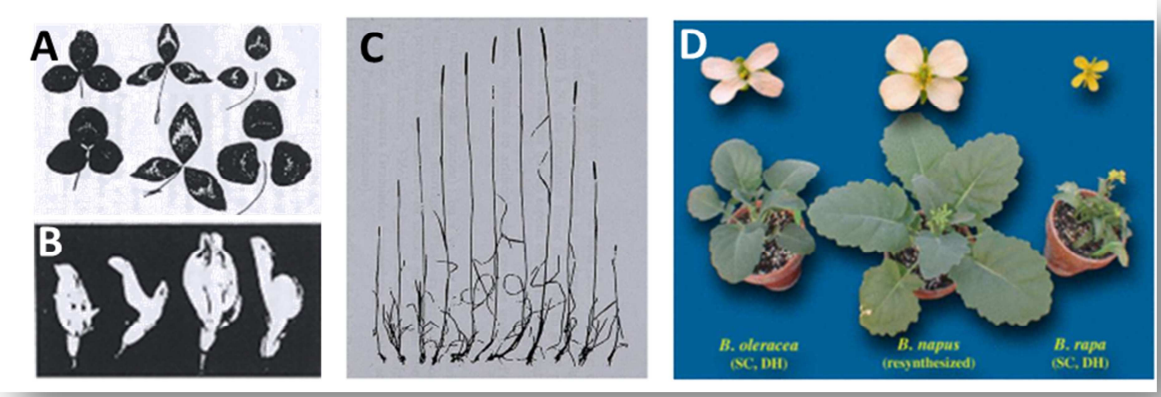


Figure 12. Illustration de l'effet 'gigantisme' de plantes polyploïdes

(A) Feuilles de trèfle violet diploïde (en haut) et tétraploïde (en bas). *Source : Elliott 1958.* (B) Fleurs de trèfle blanc diploïde (gauche) et tétraploïde (droite). *Source : Elliott 1958.* (C) Talles de Fléoles *(Phleum pratense)* de divers niveaux de ploïdie. De gauche à droite : 3x=21, 4x=28, 5x=35, 6x=42, 7x=49, 8x=56, 9x=63, 10x=70, 11x=77, 12x=84, 13x=91. *Source : Elliott 1958.* (D) Fleurs et plantes des lignées parentales *B. oleracea* à gauche (diploïde CC), *B. rapa* à droite *(*diploïde *AA*), et de *B. napus* re-synthétisé au milieu (tétraploïde AACC). *Source : J. Chris Pires.*

Chez les spartines, de nombreuses observations ont été réalisées chez différentes lignées allopolyploïdes. *Spartina anglica* (espèce dodécaploïde) présente une large amplitude écologique colonisatrice. Ses aptitudes dépassent celles de ses parents ; elle se développe sur le littoral (substrat sableux et vaseux) et peut ainsi rapidement supplanter la végétation préexistante (Ainouche *et al.* 2009 ; Chelaifa 2010). Par ses aptitudes à fixer les sédiments à l'aide d'un système racinaire puissant et de profonds rhizomes, son expansion est envahissante au niveau international. De plus, chez le genre *spartina* des hybridations sont intervenues à deux reprises entre les mêmes parents (♀S. *alterniflora* x ♂S. *maritima*) au Pays Basque et en Angleterre. Ces deux hybrides présentent une morphologie très différente (Chelaifa 2010) ; l'un étant très semblable au parent maternel, et l'autre, morphologiquement intermédiaire entre les deux parents. Ainsi, la polyploïdie impacte la biodiversité du milieu *via* le caractère adaptatif.

Chez *brassica napus,* il a été montré que la polyploïdisation impacte la réponse de la plante aux conditions environnementales et aux interactions GxE (Génotype X Environnement), notamment pour la vernalisation et la date de floraison (Schranz *et al.* 2004). Chez le blé hexaploïde établi il y a 10 000 ans, Dubcovsky et Dvorak en 2007 ont montré qu'il a surpassé les variétés de blés tétraploïdes existantes grâce à une plus forte adaptabilité. En effet, en plus d'une résistance accrue à plusieurs pathogènes, il peut se cultiver dans un large spectre de conditions environnementales, par la modification de sa réponse à la photopériode, la vernalisation, ou encore par la tolérance au froid, au sel, au pH faible et à l'aluminium (Dubcovsky et Dvorak 2007).

D'après Doyle *et al.* 2008, la polyploïdisation génère un choc génomique et les conséquences instantanées de la polyploïdisation sont diverses (*cf.* Figure 13) *via* l'augmentation du nombre de copies des loci pré-polyploïdie. En effet, le polyploïde possède plusieurs copies de chaque gène permettant un effet additif des allèles codominants, le masquage d'allèles délétères ou une augmentation de la

diversité allélique. Ce choc génomique induirait également des effets épigénétiques (Doyle *et al.* 2008, Parisod *et al.* 2009), notamment *via* des modifications de méthylation de l'ADN ou des histones et l'activation ou le silencing d'éléments transposables *via* la machinerie RNAi. Un haut niveau d'hétérosis peut être observé chez le polyploïde par rapport aux progéniteurs (Doyle *et al.* 2008). Il est reflété par une augmentation des capacités adaptatives *via* une reprogrammation de l'expression des gènes (par des phénomènes de sur/sous-expression, voire de silencing, *cf.* Figure 13) menant à de nouvelles régulations spatio-temporelles (sous-fonctionnalisation) et/ou l'apparition de nouvelles fonctions (néo-fonctionnalisation).

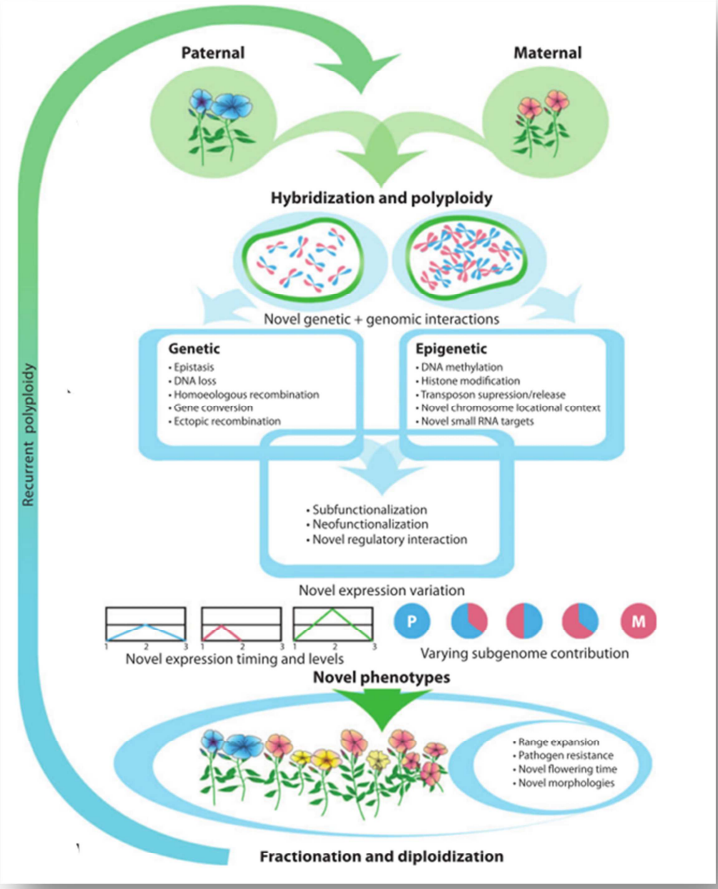


Figure 13. Processus mis en œuvre lors d'une polyploïdisation.
L'hybridation et le doublement génomique (polyploïdisation) induisent des modifications génétiques et épigénétiques. Elles génerent de nouveaux phénotypes en influant sur l'expression du génome, *via* la sous-fonctionnalisation, la néo-fonctionnalisation et les nouvelles interactions de gènes en réseaux. *Source Doyle et al. 2008.*

Doyle *et al.* 2008 résument très bien le changement dans l'expression des copies dupliquées suite à une polyploïdisation par les phénomènes de sous-fonctionnalisation et de néo-fonctionnalisation.
Dans le contexte d'une sous-fonctionnalisation où le gène ancestral avait deux fonctions (*cf.* Figure 14A), l'un des deux gènes dupliqués peut perdre l'une des fonctions, et l'autre gène peut conserver la fonction complémentaire. Dans ce contexte, les fonctions originelles ne sont plus assurées par un seul gène mais par deux protéines différentes. L'expression ou la fonction ancestrale est donc partitionnée sur les copies dupliquées post-polyploidie.

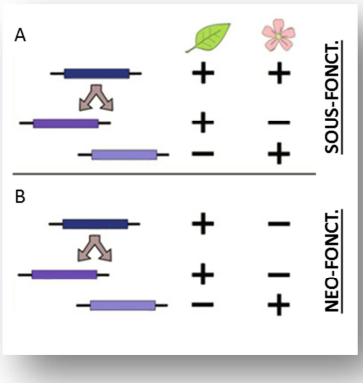


Figure 14. Illustration des modifications fonctionnelles post duplication.
A partir d'une copie ancestrale les variations observées au niveau des transcrits peuvent être de l'ordre de la sous-fonctionnalisation (A) ou de la néo-fonctionnalisation (B). Les transcrits sont modélisés par des rectangles colorés. Dans le cas de la néo-fonctionnalisation, l'expression des copies déjà préexistantes (ancestrales) sont partitionnées, selon les tissus par exemple. Dans le cas de la néo-fonctionnalisation, l'une des copies acquière des modifications structurales et fait apparaitre une nouvelle fonction (et expression) dans la fleur par exemple. *Source : Barker et al. 2011.*

Dans le contexte d'une néofonctionnalisation le gène originel code pour une fonction, et après polyploïdisation la pression de sélection peut se relâcher sur l'une de ces copies pour voir apparaître des innovations évolutives, avec une nouvelle fonction (*cf.* Figure 14A). La néo-fonctionnalisation implique qu'une des copies redondantes acquière une nouvelle capacité après avoir accumulé des mutations. La fonction ancestrale (*cf.* Figure 14B) n'est pas affectée car toujours assurée par la copie originelle.

5. Le processus de diploïdisation par dominance des sous-génomes

A l'image d'*Arabidopsis thaliana* qui a subit 3 évènements de polyploïdie mais qui est aujourd'hui diploïde (2n=10), 65% des espèces modernes sont d'anciens polyploïdes retournés à l'état diploïde (Mayrose *et al*. 2011 ; Wood *et al*. 2009). Au cours de l'évolution, les génomes polyploïdes semblent revenir à un état diploïde par perte de gènes dupliqués, ou par sous-fonctionnalisation et néo-fonctionnalisation des gènes dupliqués. En effet, malgré 3 cycles de polyploïdies, *A. thaliana* est connu pour la petite taille de son génome, issu notamment de la perte massive de gènes redondants hérités de ces WGDs (Adams *et al.* 2005, Blanc et al. 2004, Thomas *et al*. 2006). Il existe donc un mécanisme qui conduit à la suppression de gènes redondants dupliqués, appelé couramment dans la littérature 'gene partitioning' ('fractionnement' ou encore 'compartimentation' des gènes) et illustré dans la Figure 15.

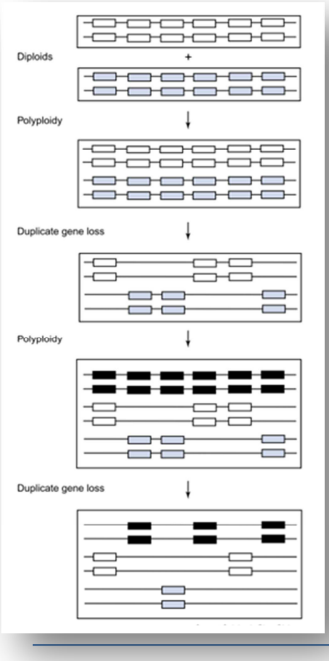


Figure 15. Illustration de la perte des gènes après polyploïdisation.
Modèle de perte massive de gènes au sein d'un génome impacté par deux WGD. Les gènes sont matérialisés par les rectangles de couleurs gris, blanc et noir. A la suite de la polyploïdie tous les gènes sont doublés. Le fractionnement est visible par la perte de gènes sur l'un des blocs dupliqués, aboutissant à la diploïdisation. Lorsqu'une deuxième polyploïdisation intervient, le même mécanisme est mis en place, épurant ainsi les gènes redondants. *Source : Adams et al. 2005.*

Au-delà de la perte des gènes, une réduction rapide de la taille du génome et du nombre de chromosomes a également été mis en évidence après polyploïdie, notamment par la perte de séquences répétées non-codantes (Renny-Byfield *et al.* 2014). La diploïdisation structurale par perte de gènes (fractionnement ou compartimentation) peut être homogène, ou se faire préférentiellement sur un des sous-génomes parentaux, post-polyploïdie. On parle dans ce cas de dominance des sous-génomes ou compartimentation asymétrique (*cf.* Figure 16). Au sein des génomes modernes, cette dominance se définit par le nombre de gènes orthologues (c'est-à-dire ancestraux) portés par les blocs chromosomiques dupliqués (ou sous-génomes). Ce phénomène de dominance des sous-génomes post-polyploïdie a été clairement observé chez différentes espèces ; l'état de l'art des connaissances sur le processus de diploïdisation par dominance des sous-génomes est présenté dans la suite de ce chapitre.

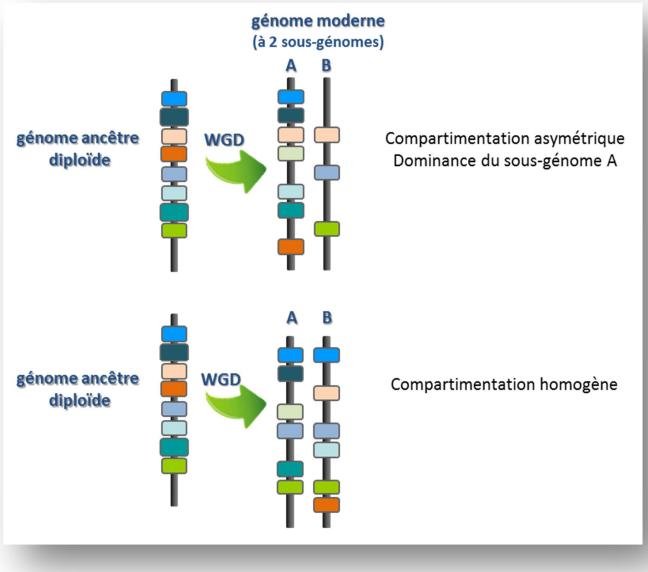


Figure 16. Représentation schématique de deux types de compartimentation (asymétrique ou homogène).
Cette figure illustre une région ancestrale (à gauche) composée de gènes (boites colorées) ayant connue un évènement de WGD pour donner naissance à deux sous-génomes (A et B). Les deux types de compartimentation sont représentés. En haut, le fractionnement est asymétrique (on parle de dominance des sous-génomes) avec une perte de gènes préférentielle sur le sous-génome B (dominance du génome A), et en bas le fractionnement est équilibré avec une perte de gènes équivalente sur les deux sous-génomes (pas de dominance). WGD ; whole genome duplication.

Chez les céréales :

Il a été démontré chez le maïs en premier lieu, que les gènes dupliqués ont été éliminés très rapidement après la polyploïdisation datant de 5 millions d'années (Freeling *et al.* 2012). Les copies ont été perdues de façon non-aléatoire à partir d'un sous-génome qui serait dit 'sensible' en faveur d'un sous-génome 'dominant' (Freeling *et al.* 2012, Murat *et al.* 2013). Ces études ont montré que la diploïdisation structurale par perte de gènes après polyploïdisation pouvait se faire préférentiellement sur un des sous-génomes parentaux post-polyploïdie (Schnable *et al.* 2011; Woodhouse *et al.* 2010).

Ce phénomène a été caractérisé plus largement chez les céréales (*cf.* Figure 17 et 18), en montrant que la dominance des sous-génomes relève d'un mécanisme ancestral (Murat *et al.* 2014). En effet, ce phénomène est conservé chez le riz, le sorgho, *Brachypodium* et le maïs, en réponse à la paléotétraploïdie datant de 90 millions d'années (Abrouk *et al.* 2012 ; Murat *et al.* 2014). Ainsi, la somme des gènes retenus de l'ancêtre est compartimentée en 2 régions ; dominante et sensible (*cf.* Figure 17). Chez toutes les céréales issues de l'ancêtre AGK à 12 chromosomes (de A1 à A 12), il est possible de classifier les chromosomes modernes (Murat *et al.* 2014 ; *cf.* Figure 18) comme provenant de sous-génomes dominants (A1/A9/A11/A4/A2/A3) ou sensibles (A5/A8/A12/A6/A7/A10).

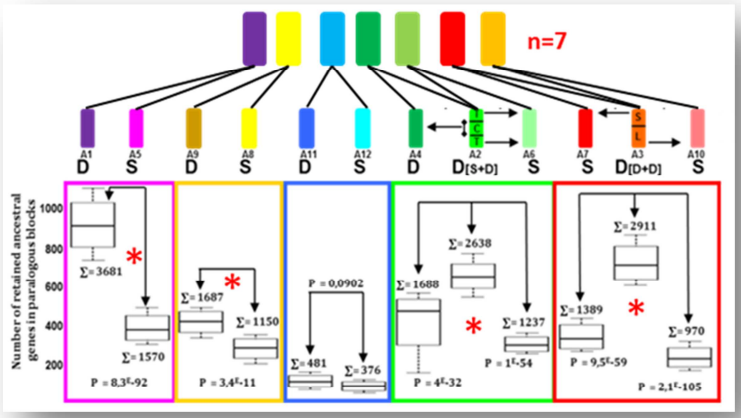


Figure 17. Dominance caractérisée chez l'ancêtre des céréales.
En haut de la figure, les chromosomes ancestraux sont représentés en couleur de A1 à A12 à partir de l'ancêtre à 7 chromosomes (ancêtre AGK avant fusion chromosomique). En bas de la figure, les effectifs de gènes orthologues retenus au sein des 12 blocks paralogues ont été comptabilisés chez le riz, le maïs, le sorgho et *Brachypodium*. Les différences significatives entre compartiments D et S sont notées par une étoile rouge. *Adapté de Murat et al. 2014.*

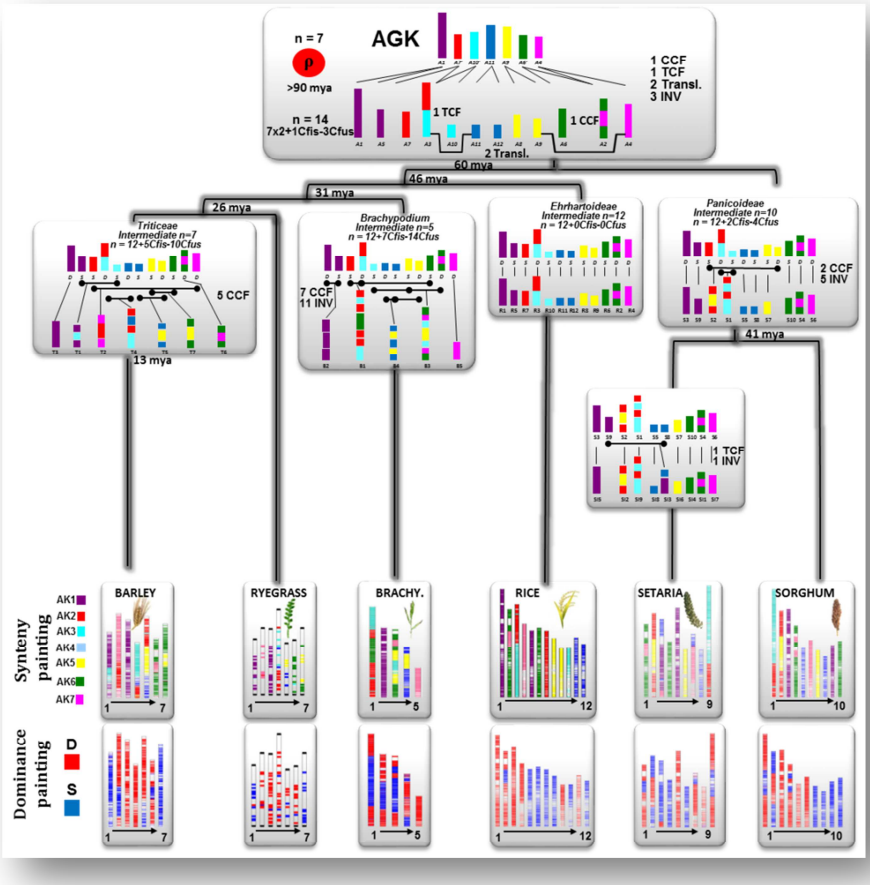


Figure 18. Compartimentation des sous-génomes dominants et sensibles chez les céréales modernes.
La synténie au sein des génomes modernes est représentée en bas (synteny painting) où les chromosomes modernes sont représentés par un code de 7 couleurs provenant des 7 chromosomes ancestraux et permettant de retracer leur origine. En bas, la compartimentation 'D et S' est représentée sur ces mêmes chromosomes (dominance painting) par un code à 2 couleurs ; rouge, dominant et bleu, sensible. Le nombre de réarrangements par fusion (centromérique CCF ou télomérique TCF), translocation, inversion est indiqué sur chaque branche d'espèce. La spéciation des différentes espèces (et ancêtres) est indiquée en millions d'années (mya). *Adaptée de Murat et al. 2014*

Ainsi, les génomes de céréales modernes, qui n'ont pas subi de duplication ou de polyploïdisation supplémentaire depuis la paléo-tétraploïdie peuvent être décomposés en 7 blocs dominants et 7 blocs sensibles. Ainsi, ces études suggèrent qu'après une duplication totale du génome, chaque chromosome ancestral a donné lieu à deux copies où le nombre des gènes retenus est significativement différent.
Parmi les gènes étudiés, les microARN suivent ce modèle ; un enrichissement de la conservation d'une des deux copies ancestrales est observé sur les chromosomes modernes de céréales (Abrouk *et al.* 2012). Cependant, même si la grande majorité des gènes répondent à ce mécanisme de conservation compartimentée post-polyploïdie, les facteurs de transcription ou plus généralement les fonctions impliquées dans la régulation des gènes (tels que les microARN) sont dits 'résistants' à la diploïdisation (Abrouk *et al.* 2012). Ces gènes sont conservés préférentiellement en doubles copies après duplication.

Chez les brassicaceae **-** La dominance des génomes a été également caractérisée suite à des évènements de duplication beaucoup plus récents, de moins de 10 millions d'années, notamment chez les *brassicaceae* (Cheng *et al.* 2012 ; Murat *et al.* 2015). *Brassica rapa,* génome diploïde entièrement séquencé (Wang *et al.* 2011), a subi une hexaploïdie il y a ~10 millions d'années (Cheng *et al.* 2012, 2013). Les trois compartiments post-hexaploïdie ont subi un fractionnement non-aléatoire donnant naissance à un sous-génome appelé LF (Less Fractioned), ayant conservé un maximum de gènes ancestraux, comparativement aux deux autres sous-génomes sensibles MF1 (Medium Fractionated) et MF2 (Most Fractionated). Ce génome alors hexaploïde, est par la suite retourné à l'état diploïde pour aboutir à la structure moderne de *Brassica rapa* à 10 chromosomes.
L'évolution du génome de *Brassica rapa* fait étrangement écho à l'évolution du génome du blé. En effet au sein de l'équipe de Jérôme Salse, l'ancêtre pré-hexaploïdie de *Brassica rapa (nommé* PCK pour *Proto-Calepineae Karyotype*) a été reconstruit en 7 chromosomes et porteur de 21 035 gènes. Cet ancêtre PCK à 7 chromosomes serait donc à l'origine des trois sous-génomes de *Brassica rapa* par deux évènements successifs d'allopolyploïdie (nommé « two-step theory » *cf.* Figure 19; Cheng *et al.* 2012).

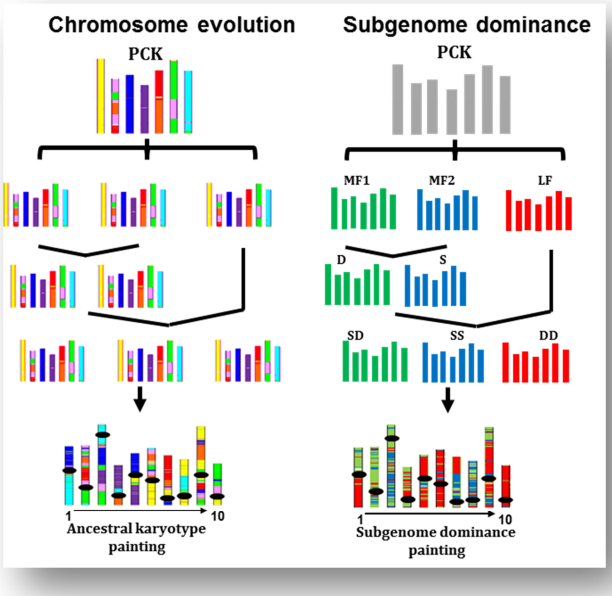


Figure 19. Représentation de la « two-step theory » chez *Brassica rapa*

L'ancêtre PCK (proto-Calepineae karyotype) est représenté en haut, à partir duquel ont divergé les trois sous-génomes LF, MF1 et MF2 du génome de *Brassica rapa*. Ainsi, *Brassica rapa* est issu de 2 cycles de polyploïdisation avec un retour à l'état diploïde donnant naissance au génome actuel composé de 10 chromosomes (en bas). En bas de la figure, le caryotype de *Brassica rapa* est représenté à gauche selon les 8 couleurs des 8 chromosomes ancestraux et à droite selon les trois couleurs de la dominance LF (« Less Fractionated », Dominant en rouge), MF1 (« Moderately Fractionated », intermédiaire en vert) et MF2 (« More Fractionated », Sensible en bleu). Les points noirs représentent les centromères. *Adapté de Murat et al. 2015.*

Selon cette théorie, les progéniteurs des sous-génomes MF1 et MF2 après hybridation ont abouti à un état tétraploïde avec 14 chromosomes. Une diploïdisation compartimentée (dominance des sous-génomes) aurait eu lieu, supprimant la redondance génique entre les deux sous-génomes préférentiellement sur MF2 amenant la dominance de MF1. Puis l'hybridation entre cet ancêtre à 14 chromosomes avec le progéniteur de LF aurait mené à un ancêtre à 21 chromosomes avec une sur-dominance de LF sur MF1 et MF2. Finalement, cette diploïdisation génique aurait donné naissance au génome actuel de *Brassica rapa* composé de 10 chromosomes et porteur de 39 498 gènes. Au-delà de la force de rétention en gènes ancestraux qui le caractérise, le génome dominant LF a été caractérisé comme globalement plus exprimé et moins polymorphe au niveau de sa diversité allélique à l'échelle populationnelle (Cheng *et al.* 2013).

Chez les polyploïdes modernes **-** Même si l'histoire évolutive du blé ressemble fortement à celle de *Brassica rapa, c*hez les polyploïdes actuels, le phénomène de dominance des sous-génomes n'a pas été clairement établi. Chez le coton par exemple, l'analyse de son génome s'est cantonnée à l'analyse de la paléo-dominance (Renny-Byfield *et al.* 2015) héritée d'une tétraploïdie datant de 1 à 2 millions d'années. Les résultats montrent que le sous-génome MF (Most Fractionated), est pauvre en gènes ancestraux, et riche en TE (et par conséquent en siRNA). L'analyse d'expression des génomes montre que MF est plus faiblement exprimé en moyenne au niveau génique, et ce, par rapport au sous-génome LF (less fractionated). Chez le colza, espèce tétraploïde depuis >7500 ans, les prémices d'une dominance ont été observés (Chalhoub *et al*. 2014) avec un fractionnement non-aléatoire dans l'effectif des gènes orthologues. Le sous-génome A possède 34 255 gènes conservés et le sous-génome C, 38 661 gènes. A l'échelle nucléotidique, les échanges partiels de gènes entre les sous-génomes représentent ~86% des différences alléliques entre *B. napus* et ses progéniteurs. Le sous-génome A montre 1,3 fois plus de conversions du génome A vers C (16 938 gènes concernés avec au moins deux sites de conversion pour le sous-génome A contre 13 429 gènes pour le sous-génome C). Ces résultats reflètent une certaine plasticité structurale en faveur du sous-génome C, cependant aucune différence n'est relatée en termes d'expression des copies homéologues.

Spartina anglica est un excellent modèle pour étudier l'impact de la polyploïdie (Parisod *et al.* 2010) sur l'organisation et la régulation des génomes, puisque cette espèce a subi un évènement de duplication datant seulement d'une centaine d'années. Cependant le niveau de ploïdie est complexe, puisque *S. anglica* est allododecaploïde, issu du doublement de parents hexaploïdes (Rousseau-Gueutin *et al.* 2016). De plus, chose assez exceptionnelle, les progéniteurs demeurant à l'état naturel et clairement identifiés, s'avèrent être les mêmes pour l'hybride homoploïde S. × *townsendii* et l'allopolyploïde *S. anglica* (*cf.* Figure 20).

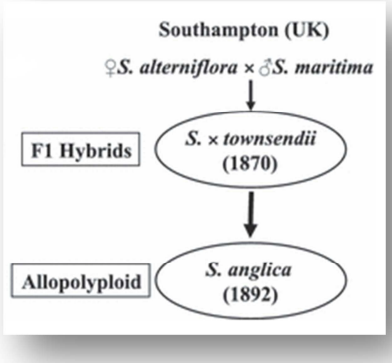


Figure 20. Origine de *Spartina anglica*
Cette figure illustre les différentes espèces de spartines existantes à l'état naturel. S. × *townsendii* est issu d'une hybridation homoploïde entre S. *alterniflora* et *S. maritima*, alors que *S. anglica* est issu d'une allopolyploïdisation entre ces mêmes parents. *Source : Parisod et al. 2010.*

Ainsi, chez les spartines (Parisod *et al.* 2010), il est possible de séparer l'effet de l'hybridation et l'effet du doublement du génome proprement dit (lors de l'allopolyploïdisation). Même s'il semble que les modifications structurales soient plus marquées chez l'hybride, elles apparaissent préférentiellement sur le sous-génome maternel S. *alterniflora* (Parisod *et al.* 2010). De même, la méthylation des TE maternels est altérée, laissant supposer une connexion possible entre l'activation des TE et la réorganisation structurale post-polyploïdie (Parisod *et al.* 2009). Dans une autre étude, un effet important mais différent, a été observé entre hybride et allopolyploïde au niveau de l'expression des gènes (Chelaifa *et al.* 2009). Une dominance maternelle est confirmée suite à l'hybridation, mais atténuée lors de la duplication chez *S. anglica.* Enfin, le polyploïde *S. anglica* montre un fort taux de gènes surexprimés (Chelaifa *et al.* 2009).

6. Le comportement méiotique diploïde du blé

La polyploïdisation est un processus tout à fait naturel de doublement du matériel génétique qui joue un rôle majeur pour l'évolution des espèces et leurs spéciations (Soltis *et al.* 2009). Lors de l'hybridation, une nouvelle espèce est produite résultant de deux génomes différents (ou plus), enveloppés dans un même noyau. Le génome global doit alors orchestrer l'expression des gènes, adapter les nouvelles interactions entre les sous-génomes, la réplication de l'ADN et assurer sa descendance. La méiose est un événement central dans le cycle de vie des cellules eucaryotes, qui assure la production de gamètes haploïdes qui, après fécondation, rétabliront l'état cellulaire initial (diploïde). La méiose diffère de la mitose par une étape de séparation des paires de chromosomes homologues, précédant la séparation des chromatides sœurs. Les chromosomes homologues s'associent en bivalents, puis ségrégent lors de la première division, tandis que les chromatides sœurs se séparent seulement lors de la deuxième division qui s'apparente elle à une mitose. L'appariement et la recombinaison des chromosomes homologues constituent un préalable requis pour leur séparation correcte. Cette étape peut être en théorie complexe chez les espèces polyploïdes du fait de la présence des chromosomes apparentés homologues et homéologues, qui peuvent présenter suffisamment de similitudes pour s'apparier et se recombiner (Cifuentes *et al*. 2009). Pourtant, un comportement méiotique régulier (formation de bivalents à la méiose) est indispensable pour assurer la ségrégation correcte des chromosomes et par conséquent la stabilité du génome et la fertilité des plantes. En effet, lorsque les chromosomes homéologues s'apparient et se recombinent, de nombreuses irrégularités méiotiques sont obtenues. Elles aboutissent entre autre, à la production de gamètes aneuploïdes (Sanchez-Moran *et al*. 2001) et/ou de gamètes remaniés, qui ne sont généralement pas viables, entrainant une baisse de la fertilité (Ramsey & Schemske, 2002). Alors que chez les diploïdes la formation de bivalents est stricte, chez les autotétraploïdes et allopolyploïdes des comportements méiotiques trivalents ou tétravalents, sont observés bien qu'en faibles proportions (*cf.* Figure 21).

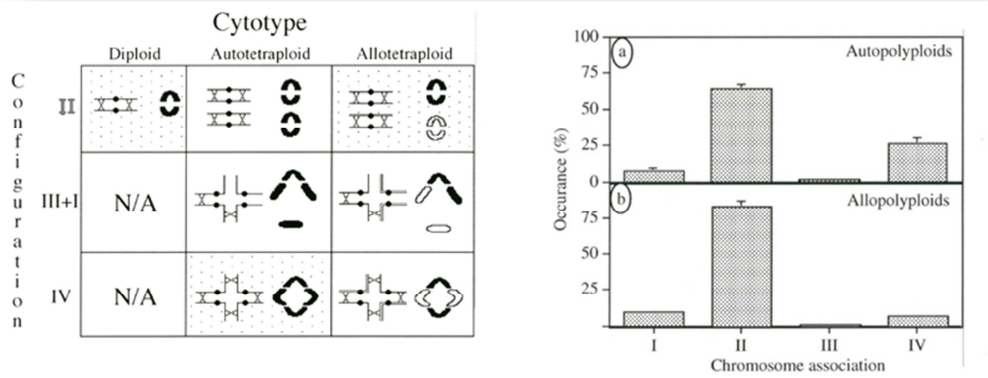


Figure 21 : Comportement méiotique chez les espèces diploïdes et polyploïdes

La figure de gauche illustre les différentes configurations observables lors de la méiose chez le diploïde, l'autotétraploïde et l'allotétraploïde. Les associations chromosomiques retrouvées sont de trois types ; bivalents (II) ; trivalents et univalents (III+I) ; tétravalents (IV). La figure de droite représente les proportions retrouvées de chaque configuration (occurrences en %) (I à IV) chez l'autopolyploïde (a) et l'allopolyploïde (b). *Source : http://www.vivelessvt.com/*

Cependant, les espèces polyploïdes de blé ont stabilisé leur comportement méiotique pour obtenir un appariement strictement diploïde, c'est-à-dire une formation de bivalents exclusivement entre chromosomes homologues. En effet, malgré le contexte chromosomique complexe des blés hexaploïdes et tétraploïdes, les cytogénéticiens ont observé qu'ils présentent tous un comportement régulier diploïde à la méiose (Feldman 1994, Griffiths *et al.* 2006). Le comportement régulier au niveau cytologique chez le blé est assuré génétiquement par un réseau multigénique dont le gène Ph1 (*pairing homeologous*) a un effet majeur (Cifuentes *et al.* 2009). Ph1 peut bloquer l'apparentement des chromosomes homéologues chez le blé polyploïde pour permettre aux chromosomes homologues de se coupler. Ainsi, ce gène limite l'aneuploïdie et favorise donc le maintien du comportement méiotique régulier pour assurer la fertilité de l'espèce. L'organisation spécifique de Ph1 sur le chromosome 5B a été observée chez tous les blés polyploïdes, mais n'est présente dans aucune espèce diploïde des genres *Triticum* et *Aegilops.* Par contre, elle est présente chez T. *timopheveii* (2n=4x=28, AAGG), un autre blé allotétraploïde naturel. Ceci suggère que le locus Ph1 est apparu peu après l'évènement de polyploïdisation et qu'il a été maintenu dans différents allopolyploïdes naturels de blé (Griffiths *et al.* 2006).

7. Questionnement scientifique de la thèse

Les travaux récents effectués au sein de l'équipe ont clairement caractérisé l'histoire évolutive des génomes de céréales à partir d'un ancêtre commun il y a 90 MYA. Ils ont établi que la polyploïdie, induisant l'apport de copies surnuméraires de chaque gène, est un évènement récurrent de l'évolution des plantes. Il apparaît acquis que la polyploïdie constitue un 'choc génomique' qui permet une 'reprogrammation' des génomes post-polyploïdie. La polyploïdie est suivie d'une diploïdisation structurale (perte des gènes) et fonctionnelle (modification de l'expression des gènes) conduisant au phénomène de dominance des sous-génomes (dérivant des compartiments dominants et sensibles). Ces

phénomènes ont été notés sur des espèces modernes ayant subi des polyploïdisations anciennes, de telle sorte que ces espèces sont aujourd'hui diploïdes.

Cette dominance des sous-génomes existe-t-elle chez les polyploïdes modernes ? Le blé hexaploïde (AABBDD) constitue dans ce contexte scientifique un excellent modèle permettant l'étude de l'impact de la polyploïdie sur la structure et la fonction des gènes, car celui-ci a subi des évènements anciens et récents de polyploïdie, il y a 90 MYA et jusqu'à 10 000 ans (*i.e.* paléo et néo-polyploïdie). Même si la communauté scientifique ne dispose pas encore des 21 pseudomolécules exhaustives, de nombreuses ressources génomiques sont disponibles permettant l'étude de l'impact de la polyploïdie sur l'organisation et la régulation du génome du blé tendre moderne.

Quel est le niveau de redondance structurale (au travers du contenu en gènes, Chapitre I) et fonctionnelle (au travers de l'expression des gènes, Chapitre II) des sous-génomes du blé tendre issus d'événements anciens et récents de polyploïdie ? Quel est l'impact de l'asymétrie structurale et fonctionnelle des sous-génomes sur l'élaboration de phénotypes (au travers du caractère de tallage, Chapitre III) travaillés en sélection ? Ces questions seront traitées dans le cadre des chapitres suivants au travers de 3 articles scientifiques dont je suis le premier auteur.

CHAPITRE 1

Impact de la polyploïdie sur l'organisation du génome du blé tendre

1. Introduction à l'étude de l'impact de la polyploïdie sur l'organisation du génome du blé tendre

1.1. Origine paléo- et néo-polyploïde du génome du blé tendre

Comme il a été discuté dans la partie introductive (*cf.* section « l'évolution du génome du blé tendre hexaploïde », page 9), le blé tendre moderne est le fruit d'une histoire évolutive complexe. Il est issu d'un ancêtre AGK paléo-tétrapolyploïde commun aux céréales (datant de ≈90 MYA), à 12 chromosomes. Il a subi 5 fusions chromosomiques, 2 translocations, suivi de 2 événements de polyploïdisation récents (allopolyploïdisation). Le premier événement a eu lieu il y a environ 0,5 MYA donnant naissance au tétraploïde *Triticum turgidum* (génome AB) et plus récemment, il y a 10 000 ans, pour former l'espèce *Triticum aestivum* (génome hexaploïde ABD), l'actuel blé tendre. Ainsi, pour une région ancestrale des céréales, 6 régions dupliquées sont potentiellement présentes chez le blé ; deux régions issues de la paléo-tétraploïdie (2 paléo-sous-génomes ohnologues) et 3 régions issues de la néo-hexaploïdie (3 néo-sous-génomes homéologues ; *cf.* Figure 22).

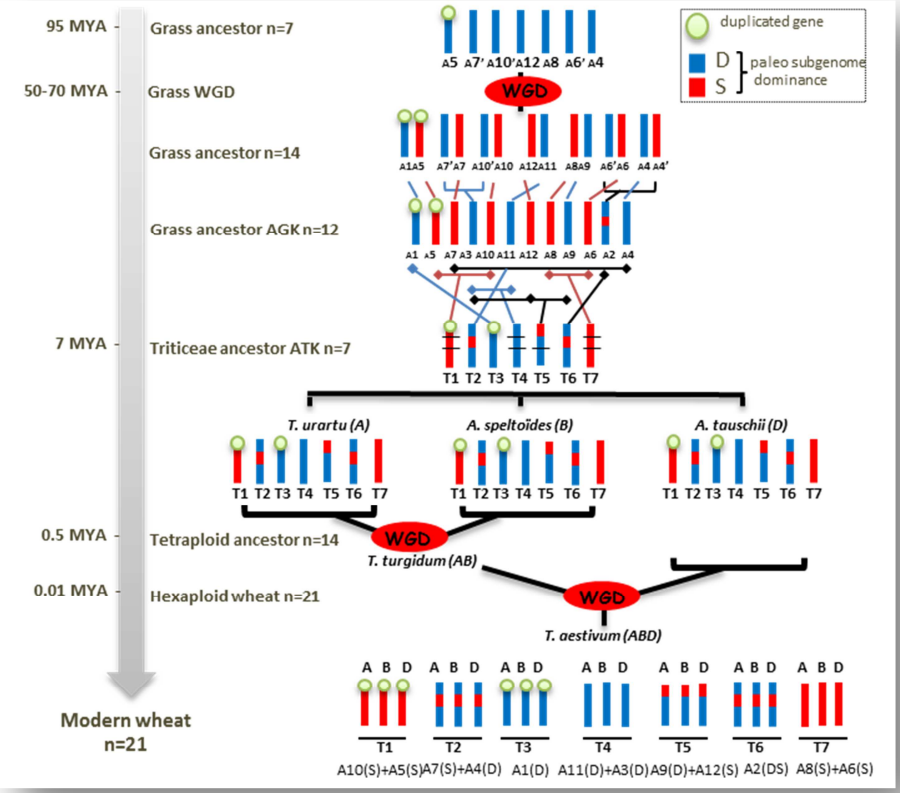


Figure 22. Modèle évolutif du génome du blé tendre moderne à partir de l'ancêtre des céréales de 7 et 12 chromosomes.

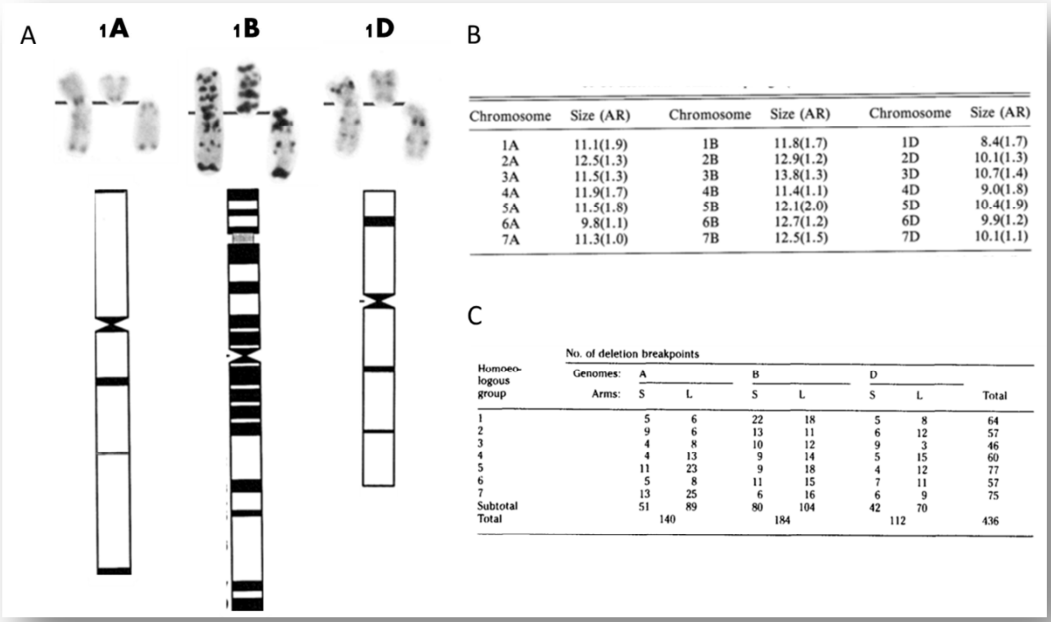
L'histoire évolutive du blé est schématisée dans cette figure à partir de l'ancêtre des céréales à 7 chromosomes. Cet ancêtre a subi une duplication totale de son génome (WGD), doublant le contenu chromosomique, puis deux réarrangements pour aboutir à l'ancêtre AGK à 12 chromosomes. Cette WGD a engendré un effet de dominance entre les deux paléo-sous-génomes, matérialisée ici en bleu, pour la fraction dominante (A1-2-3-4-9-11) et en rouge, pour la fraction sensible (A5-6-7-8-10-12). Cette compartimentation peut être identifiée au sein du

génome blé moderne composé de 21 chromosomes en tenant compte des 5 fusions chromosomiques (ayant donné naissance aux 7 groupes chromosomiques des *Triticeae*). A titre d'exemple, un gène ancestral matérialisé par un cercle vert, présent en une copie dans l'ancêtre est potentiellement présent en 6 copies chez le blé moderne (2 paléo-sous-génomes x 3 néo-sous-génomes). Les relations d'orthologie entre les 7 chromosomes ancestraux (A1 à A12) et les 21 chromosomes du blé tendre (T1 à T7 avec les trois sous-génomes A, B et D) sont mentionnées au bas de la figure. Ainsi la dominance et sensibilité des sous-génomes du blé tendre apparaît comme suit T1(S)=A10(S)+A5(S), T2(D/S)=A7(S)+A4(D), T3(D)=A1(D), T4(D)=A11(D)+A3(D), T5(D/S)=A9(D)+A12(S), T6(D/S)=A2(D/S), T7(S)=A8(S)+A6(S). La datation des évènements de spéciation et de duplication est indiquée en millions d'années (MYA). AGK : ancestral grass karyotype ; ATK : ancestral *Triticeae* karyotype ; WGD : whole genome duplication.

Quelles différences structurales sont observables au sein de ces six compartiments de sous-génomes du blé tendre ? Les paragraphes suivants sont le bilan de l'état de l'art de nos connaissances avant la parution de mon article sur l'asymétrie structurale observée entre les sous-génomes du blé tendre. Cet état de l'art fait le bilan des connaissances acquises sur l'asymétrie structurale observée entre les trois sous-génomes issus de la néo-hexaploïdisation (*i.e* entre A, B et D), mais non héritée de la paléo-tétraploïdisation (*i.e* entre les blocs rouges (S) et bleus (D) de la Figure 22) qui n'était pas connue avant nos travaux présentés en détails dans l'article scientifique discuté dans ce chapitre.

1.2. Asymétrie caryotypique des sous-génomes du blé tendre

Dès les premières analyses caryotypiques, les cytogénéticiens ont remarqué des différences structurales entre les génomes A, B et D du blé hexaploïde (*cf.* Figure 23A-C). En effet, concernant la taille des chromosomes, le sous-génome B semble être de plus grande taille à l'inverse du génome D, le plus petit (*cf.* Figure 23B ; Gill *et al.*1991 ; Endo 1996). Une asymétrie de taille est donc observée, avec une graduation B>A>D, qui peut être également corrélée avec le taux de breakpoint comptabilisé à partir des lignées de délétions de Sears (*cf.* Figure 23C ; B#184 > A#140 > D#112 ; Gill *et al.* 1991 ; Endo 1996). En effet, ces lignées, issues toujours du cultivar *Chinese Spring*, ont subi des cassures structurales (pertes de fragments ou bras de chromosomes) qui peuvent être localisées à l'aide de marqueurs moléculaires (non amplifiés si la région qui les porte a été perdue). La région manquante, est alors appelée bin de délétion. Il a été montré que la plupart des points chauds de ruptures (breakpoint) ont lieu à la jonction entre hétérochromatine et euchromatine révélées par C-Banding (Endo 1996). Cette méthode révèle les zones d'hétérochromatine contenant de l'ADN satellite constitué de séquences courtes et répétées en tandem. Là encore, une asymétrie est clairement observée avec une abondance de 'C-Bandes' sur le sous-génome B, par rapport aux deux autres sous-génomes (*cf.* Figure 23A). Cette différence de structure chromatinienne des sous-génomes soulève un grand nombre de questions quant à son impact sur l'organisation et la régulation des gènes, puisque l'on sait que l'hétérochromatine condensée est transcriptionnellement inactive et pauvre en gènes. Chez le riz, comme chez beaucoup d'espèces, la distribution d'hétérochromatine est répartie dans les régions péricentromèriques, mais elle n'est pas homogène entre les chromosomes (Cheng *et al*. 2001). L'hétérochromatine varie donc selon les chromosomes, mais présente aussi un faible taux de recombinaison et de teneur en gènes (chez le sorgho ; 29 gènes par Mbp d'hétérochromatine et 81 gènes par Mbp pour l'euchromatine ; Kim *et al.* 2005).



Chromosome	Size (AR)	Chromosome	Size (AR)	Chromosome	Size (AR)
1A	11.1(1.9)	1B	11.8(1.7)	1D	8.4(1.7)
2A	12.5(1.3)	2B	12.9(1.2)	2D	10.1(1.3)
3A	11.5(1.3)	3B	13.8(1.3)	3D	10.7(1.4)
4A	11.9(1.7)	4B	11.4(1.1)	4D	9.0(1.8)
5A	11.5(1.8)	5B	12.1(2.0)	5D	10.4(1.9)
6A	9.8(1.1)	6B	12.7(1.2)	6D	9.9(1.2)
7A	11.3(1.0)	7B	12.5(1.5)	7D	10.1(1.1)

Homoeologous group	No. of deletion breakpoints: Genomes: A		B		D		
	Arms: S	L	S	L	S	L	Total
1	5	6	22	18	5	8	64
2	9	6	13	11	6	12	57
3	4	8	10	12	9	3	46
4	4	13	9	14	5	15	60
5	11	23	9	18	4	12	77
6	5	8	11	15	7	11	57
7	13	25	6	16	6	9	75
Subtotal	51	89	80	104	42	70	
Total	140		184		112		436

Figure 23. Illustration de l'asymétrie caryotypique des sous-génomes du blé tendre.
Idiogramme C- Banding du groupe chromosomique 1. Les 'C-Bandes' observées sont reportées en noir sur le schéma des chromosomes. *Source : Gill et al. 1991*. (B) Taille des 21 chromosomes et ratio de la longueur des deux bras, *Souce : Gill et al. 1991*. (C) Distribution des breakpoints sur les bras courts et longs des 21 chromosomes du blé tendre. *Source : Endo 1996*. S : bras courts ; L : bras longs ; AR : arm ratio.

1.3. Asymétrie génique des sous-génomes du blé tendre

Une première étude d'Akhunov *et al.* 2003, cartographiant un lot de 845 gènes, a noté un biais du contenu en gènes avec respectivement 746, 686 et 593 gènes pour les sous-génomes D, B et A. Puis, une analyse de la distribution des gènes à grande échelle a été menée par Qi *et al.* 2004, qui, par hybridation de sondes sur les lignées de délétions de Sears (Endo *et al.* 1996), a permis de localiser précisément sur les 21 chromosomes du blé 5 762 EST (*expressed sequence tag*). Ces gènes ont été assignés à 16 099 loci ; soit une moyenne de redondance de 2,8 positions de cartographie par marqueurs. Les résultats montrent que 42 % des sondes sont cartographiées à 5 loci, 46 % à 3 loci (potentiellement les 3 homéologues), et 12 % sur 1 à 2 loci (correspondant à la perte d'au moins une des trois copies homéologues). Une asymétrie de l'organisation structurale des gènes a été observée dans cette étude entre les 3 sous-génomes, tant en nombre qu'en densité de gènes (*cf.* Figure 24A-B). Parmi les 16 099 loci assignés, 2 048 loci ont été localisés à un seul chromosome homéologue. La distribution de ces singletons montre que les sous-génomes A et D sont relativement homogènes en nombre (5 146 pour le A et 5 179 pour le D) de gènes, mais que le génome B comporte 11% d'EST assignées de plus que les deux autres sous-génomes (5 774). Lorsque la taille physique des chromosomes est prise en compte (avec l'ordre croissant D<A<B), le génome D montre une plus forte densité en gènes par rapport aux deux autres sous-génomes (*cf.* Figure 24B). Ces résultats suggèrent une organisation des gènes non homogène au sein des sous-génomes homéologues, bien que dérivant d'un ancêtre commun ATK. Enfin, l'assignation précise des EST dans les bins de délétion a montré que les régions télomériques étaient plus denses en transcrits que celles proches des centromères (*cf.* Figure 24A).

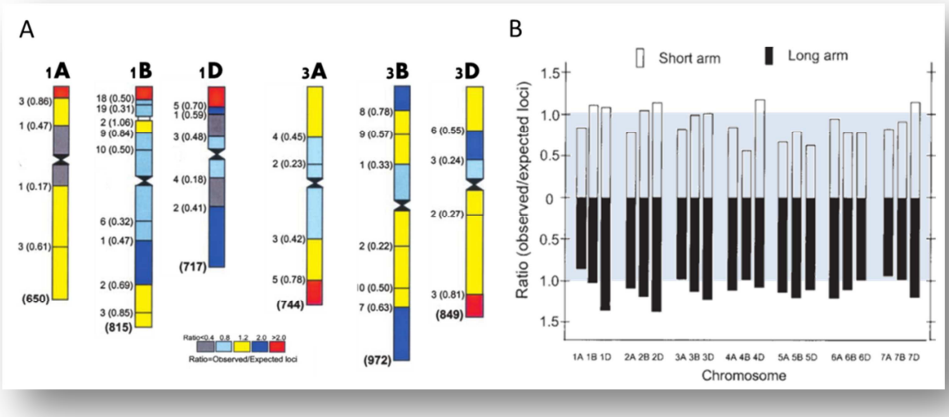


Figure 24. Illustration de l'asymétrie structurale des sous-génomes du blé tendre.
(A) Densité en gènes, matérialisée en couleurs sur les chromosomes 1 et 3, *via* le ratio observé/attendu. Le rouge matérialise la plus forte densité (échelle en bas). La taille des chromosomes est indiquée entre parenthèses ainsi que le nombre de gènes total cartographié (en gras). (B) Densité en gènes sur les bras longs et courts des 21 chromosomes selon le ratio observé/attendu prenant en compte la taille des chromosomes. *Source : Qi et al. 2004.*

Cette asymétrie du contenu en gènes des sous-génomes a également été observée grâce au séquençage de clones BACs. A titre d'exemple, le locus *Ha* (*Hardiness* ; Gu *et al.* 2006), impliqué dans la dureté du grain de blé par les gènes *Pina* et *Pinb* (codant pour la friabiline, responsable du caractère 'soft' du grain de blé), a été étudié en 2005. La banque BAC de la variété référence *Chinese Spring* a été construite par l'INRA (1 138 944 clones, d'une taille moyenne de 130.0 Kb) et l'analyse des séquences au locus *Ha* montre une asymétrie structurale (*cf.* Figure 25). En effet, l'analyse des BACs au locus *Ha* montre la présence des gènes *Pin* sur le sous-génome D, alors qu'ils sont absents des sous-génomes A et B de l'hexaploïde. Ce résultat a permis aux auteurs de proposer un scénario évolutif où ce caractère ancestral a été perdu chez le blé tétraploïde par la délétion indépendante des deux gènes *Pin*, et a été réintroduit chez le blé tendre *via* le génome D lors du dernier évènement de polyploïdisation. Cet exemple illustre à l'échelle d'un locus chez le blé hexaploïde, l'asymétrie des sous-génomes dans leur contenu en gènes et l'impact sur le phénotype (ici la dureté du grain).

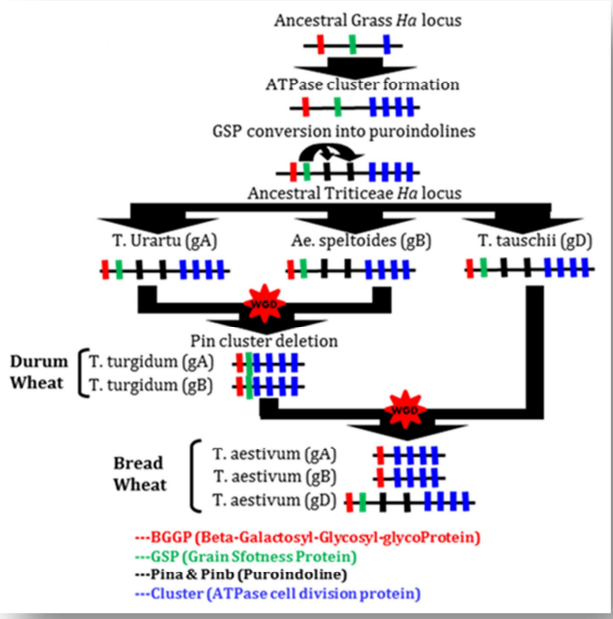


Figure 25. Modèle évolutif du locus *Ha* chez le blé tendre.
Le modèle évolutif du locus *Ha* est schématisé dans cette figure à partir de l'ancêtre des céréales (en haut), aboutissant au blé tendre hexaploïde (en bas). Cet ancêtre a connu une duplication segmentale multipliant les gènes du locus représentés par les rectangles, dont les copies *Pina* et *Pinb* en noir (*cf.* légende au bas). Ces gènes ont été perdus après la tétraploïdisation sur les génomes A (gA) et B (gB), mais réintroduits lors de l'hybridation avec le génome D g(D). Les gènes *Pina* et *Pinb* codent pour la friabiline responsable du caractère 'soft' du blé tendre. WGD : whole genome duplication. *Source : Chantret et al. 2005.*

Conclusion : la littérature antérieure à mes travaux de thèse reporte une asymétrie structurale entre les sous-génomes du blé tendre hexaploïde par des approches peu résolutives de cytogénétique, de cartographie génétique ou de séquençage de clones BAC. Dans l'article qui suit, paru dans la revue 'Plant Journal' en décembre 2013, j'ai étudié la plasticité structurale des sous-génomes du blé pour proposer le premier modèle évolutif introduisant la dominance des sous-génomes post-polyploïdie comme moteur de l'asymétrie des sous-génomes chez le blé hexaploïde moderne.

Cet article, dont je suis le premier auteur, rassemble 18 cosignataires dont il parait important de définir le rôle et la contribution dans les résultats présentés. Ainsi, dans le cadre de cet article, j'ai intégré les données de la littérature, puis j'ai défini et caractérisé 5 234 *marqueurs COS sur la base des données de synténie de l'équipe. J'ai coordonné le séquençage de ces marqueurs sur 5 variétés de blé hexaploïde et 3 variétés de blé tétraploïde par le prestataire GATC. Y. Bidet m'a permis d'accéder à la plateforme de séquençage GS454 pour réaliser des runs de séquençage préliminaires. A la réception des séquences, j'ai collaboré avec F. Murat (Ingénieur bioinformaticien) pour l'analyse des séquences et j'ai encadré S. Foucrier (Ingénieur en biologie moléculaire) pour la cartographie des SNP qui a permis la construction de la carte génétique consensus WCGM 2013 par U. Masood Quraishi (Doctorant). A partir de ces données, j'ai réalisé l'analyse des sous-génomes du blé tendre hexaploïde comme décrit dans l'article. Enfin, S. Guizard (Ingénieur bioinformaticien) avec les membres de l'URGI (Raphael Flores, Delphine Steinbach, Michael Alaux, Hadi Quesneville) ont développé l'interface Syntenyviewer. Les autres auteurs (Jaroslav Deloze, Tzion Fahima, Hikmet Budak, Beat Keller, Silvio Salvi, Marco Maccaferri, Catherine Feuillet) ont participé au financement de ces travaux dans le cadre du projet européen Triticeae Genome.*

2. Article paru dans la revue 'Plant Journal' en 2013

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) doi: 10.1111/tpj.12366

Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes

Caroline Pont[1,†], Florent Murat[1,†], Sébastien Guizard[1], Raphael Flores[2], Séverine Foucrier[1], Yannick Bidet[3], Umar Masood Quraishi[1], Michael Alaux[2], Jaroslav Doležel[4], Tzion Fahima[5], Hikmet Budak[6], Beat Keller[7], Silvio Salvi[8], Marco Maccaferri[8], Delphine Steinbach[2], Catherine Feuillet[1], Hadi Quesneville[2] and Jérôme Salse[1,*]

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **76,** 1030–1044 doi: 10.1111/tpj.12366

Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes

Caroline Pont[1,†], Florent Murat[1,†], Sébastien Guizard[1], Raphael Flores[2], Séverine Foucrier[1], Yannick Bidet[3], Umar Masood Quraishi[1], Michael Alaux[2], Jaroslav Doležel[4], Tzion Fahima[5], Hikmet Budak[6], Beat Keller[7], Silvio Salvi[8], Marco Maccaferri[8], Delphine Steinbach[2], Catherine Feuillet[1], Hadi Quesneville[2] and Jérôme Salse[1,*]

[1]*INRA/UBP UMR 1095, Centre de Clermont Ferrand-Theix, 5 Chemin de Beaulieu, 63100 Clermont Ferrand, France,*
[2]*INRA/URGI, Centre de Versailles, bâtiment 18, route de Saint Cyr, 78026 Versailles cedex, France,*
[3]*Clermont Université/Plateforme GINA, Centre Jean Perrin, 58 rue Montalembert, 63011 Clermont Ferrand, France,*
[4]*Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic,*
[5]*Department of Evolutionary and Environmental Biology, Faculty of Natural Sciences University of Haifa Mt. Carmel, University of Haifa, Haifa, 31905 Israel,*
[6]*Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli, Tuzla-Istanbul, Turkey,*
[7]*Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland, and*
[8]*University of Bologna, DiSTA – Agronomy, Viale Fanin, 44, 40127 Bologna, Italy*

Received 27 August 2013; revised 1 October 2013; accepted 8 October 2013; published online 26 October 2013.
*For correspondence (e-mail jsalse@clermont.inra.fr).
†These authors contributed equally to the work.

SUMMARY

Bread wheat derives from a grass ancestor structured in seven protochromosomes followed by a paleotetraploidization to reach a 12 chromosomes intermediate and a neohexaploidization (involving subgenomes A, B and D) event that finally shaped the 21 modern chromosomes. Insights into wheat syntenome in sequencing conserved orthologous set (COS) genes unravelled differences in genomic structure (such as gene conservation and diversity) and genetical landscape (such as recombination pattern) between ancestral as well as recent duplicated blocks. Contrasted evolutionary plasticity is observed where the B subgenome appears more sensitive (i.e. plastic) in contrast to A as dominant (i.e. stable) in response to the neotetraploidization and D subgenome as supra-dominant (i.e. pivotal) in response to the neohexaploidization event. Finally, the wheat syntenome, delivered through a public web interface PlantSyntenyViewer at http://urgi. versailles.inra.fr/synteny-wheat, can be considered as a guide for accelerated dissection of major agronomical traits in wheat.

Keywords: paleogenomics, dominance, partitioning, conserved orthologous set, single nucleotide polymorphism.

INTRODUCTION

Recent comparative genomics studies based on monocot genome sequences, including Panicoideae (sorghum, Paterson *et al.*, 2009; maize, Schnable *et al.*, 2009), Ehrhartoideae (rice, IRGSP, 2005), and Pooideae (*Brachypodium*, IBI, 2010), suggest that grasses derive from $n = 7$ (alternative scenario with $n = 5$) to 12 ancestral karyotypes (named AGK for Ancestral Grass Karyotypes). Modern grass genomes were then shaped from this AGK through a shared whole-genome duplication (WGD) followed by ancestral chromosome fusion (CF) events (for review Salse, 2012). Polyploidization has been shown to be followed by genome-wide diploidization (also referenced as partitioning) through differential elimination of duplicated gene redundancy at the whole-genome level (Wang *et al.*, 2005; Freeling *et al.*, 2012; Schnable *et al.*, 2012a,b), leading to dominant (i.e. D, stable genomic compartment associated with low duplicated gene loss and reduced gene diversity) and sensitive (i.e. S, plastic genomic compartment associated with high duplicated gene loss and increased gene diversity) subgenomes in modern grass

species (Woodhouse *et al.*, 2010; Abrouk *et al.*, 2012; Schnable *et al.*, 2012a,b).

Bread wheat is a good plant model to study the impact of distinct rounds of WGD on the subgenome dominance phenomenon, as its genome comprises: (i) seven ancestral paleoduplicated blocks corresponding to the shared paleotetraploidization event identified in all known cereal genomes and dating back to 65 million years ago (hereafter, mya); as well as (ii) two recent neopolyploidization events leading to *Triticum aestivum*, which originated from two hybridizations between *T. urartu* (A genome) and an *Aegilops. Speltoides* related species (B genome) 1.5 mya, forming *T. turgidum ssp. dicoccoides*; and between *T. turgidum ssp. durum* (genomes A–B) and *A. tauschii* (D genome) 10 000 years ago (Feldman *et al.*, 1995). Wheat diploidization (between A, B and D subgenomes) has been partially investigated at the whole-genome and gene levels where recent transcriptome analysis in hexaploid wheat suggested that 39 up to 46% of the wheat homoeologous genes may have either been lost or neo-/subfunctionalized within 1.5 my of evolution (Pont *et al.*, 2011). Such structural or functional diploidization brings unbalanced polyploidy gene systems into diploid-like mode of expression (Edger and Pires, 2009).

The recent access to a large bread wheat genomic resources offers the opportunity to study in the same analysis not only the structural plasticity of paleoduplicated genes (during the last 50–70 million years of evolution) but also neoduplicated genes (during 0.1–1.5 million years of evolution) by comparing the conservation of A, B and D homoeologous gene copies, i.e. wheat homoeoalleles. Wheat genomics resources have been recently published with the release of the wheat genome shotgun sequences in hexaploid (Brenchley *et al.*, 2012) and diploids (D genome ancestor sequence in Jia *et al.*, 2013 and Luo *et al.*, 2013 as well as the A genome progenitor sequence in Ling *et al.*, 2013). However, these studies reported incomplete gene repertoire (i.e. whole-genome short reads assembled into gene models instead of whole-chromosome pseudomolecules) either not ordered or partially ordered based on synteny relationships, mainly with *Brachypodium*. Independently, genome-wide diversity maps have been also made recently available in hexaploids (Chao *et al.*, 2009; Allen *et al.*, 2011, 2013; Lai *et al.*, 2012; Winfield *et al.*, 2012; Cavanagh *et al.*, 2013), tetraploids (Saintenac *et al.*, 2011; Trebbi *et al.*, 2011; Ren *et al.*, 2013) or diploid progenitors (You *et al.*, 2011; Wang *et al.*, 2013). However, most of these studies relied on transcriptome (*i.e.* exome) sequencing from different genotypes and then are dependent on both the tissues used and the expression of the three homoeologs on the considered tested tissues and developmental conditions, then potentially leading to possible bias in homoeoallele diversity estimation due to gene silencing in the tested biological conditions.

In contrast, the evolution of the homoeologous gene space in hexaploid wheat can be investigated through comparative genomics approaches trying to introduce no (or reduced) bias in gene loss and sequence polymorphism assessment (Thomas *et al.*, 2006; Bekaert *et al.*, 2011). Up to now very few genes are physically and genetically mapped in hexaploid bread wheat. The integration of both genomic and genetic resources, described in the previous section, offers the opportunity to provide the most accurate wheat syntenic (or also referenced as computed, Pont *et al.*, 2011) gene order (i.e. defined as syntenome) and test its accuracy as this will probably represent in a long-term the wheat reference genome, until complete pseudomolecules will be publicly released for the 21 chromosomes. Such validated wheat syntenome will also offer the opportunity to perform a comprehensive analysis of the wheat gene space evolutionary plasticity during the last 100 million years and to investigate the impact of paleo- and neopolyploidization events on genome rearrangement and gene diversity. Despite insights into wheat subgenome evolution, the wheat syntenome can also be consider as an applied tool for the dissection of major traits in wheat.

RESULTS

Wheat syntenome defines paleo/neodominant and sensitive blocks

Grasses has been proposed to derive from a $n = 7$ ancestor that has been duplicated to reach a $n = 14$ intermediate followed by two chromosomal fusions to reach a $n = 12$ ancestor (Figure 1a centre of the circle and inner circle A1–A12) founder of all the modern grasses (Salse, 2012). Comparing grass genome sequences, we identified a 17 317 conserved orthologous set (COS; also referenced as protogenes) located on 12 syntenic blocks (i.e. conserved ancestral regions; CARs) covering the rice, maize, sorghum, *Brachypodium* genomes (Figure 1a inner cereal circles and Table 1) and refining the previously reported conserved gene repertoire in grasses of 9731 protogenes (Murat *et al.*, 2010). The 17 317 COS has been used as a matrix to produce a wheat syntenome where on average 2474 COS have been ordered on the seven wheat chromosome groups in respect to the position of their orthologous counterparts following the ordering priority of rice > *Brachypodium* > sorghum (*cf* wheat consensus circle on Figure 1a and Table 1). We made this wheat syntenome available through a public web interface named PlantSyntenyViewer at http://urgi.versailles.inra.fr/synteny-wheat (illustrated in Figure 1b and with raw data available in Table S1). The wheat syntenome offered the opportunity to unravel the evolutionary fate of such COS in wheat in characterizing retained *vs.* lost duplicates in response to both ancestral shared (~65 mya) and recent specific (<1.5 mya) polyploidization events. To do this, we

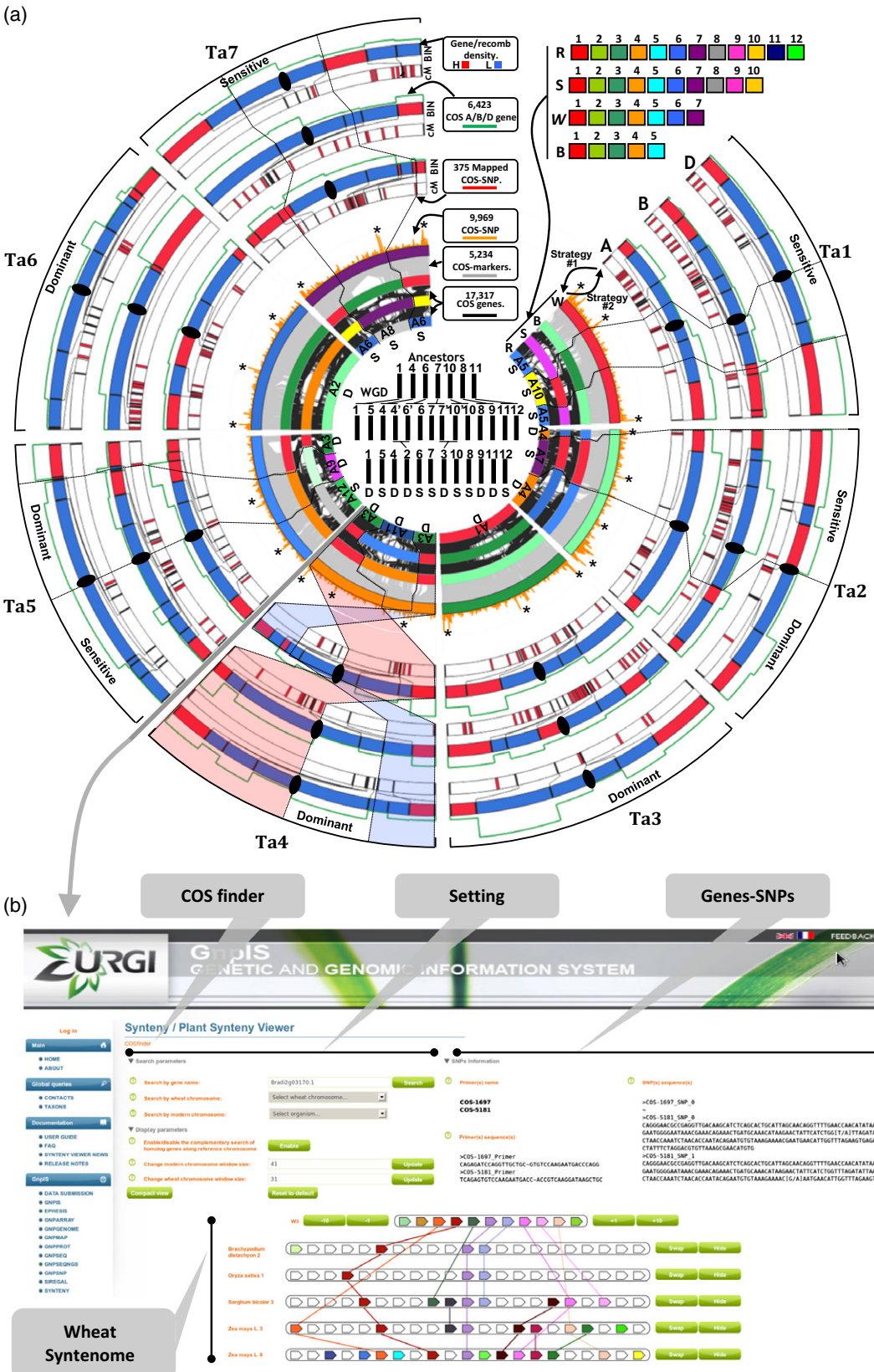
Figure 1. Wheat syntenome circles.
(a) The centre of the circle illustrates the wheat chromosomes origin from a grass ancestor structured in seven protochromosomes followed by a paleotetraploidization to reach a 12 chromosomes of known dominance (D) and sensitivity (S) blocks. Integration of the seven bread wheat chromosome groups (wheat consensus circle, 'wheat cons') with the rice (12 chromosomes, 'R' circle), sorghum (10 chromosomes, 'S' circle) and *Brachypodium* (five chromosomes, 'B' circle) chromosomes is illustrated as inner circles with 17 317 COS (black connecting lines between inner circles), 5234 COS markers selected for sequencing (grey connecting lines between inner circles) and 9969 SNPs (shown as orange distribution bars on the 'wheat cons' circle) according to their orthologous positions on rice, sorghum and *Brachypodium* chromosomes. The three external concentric circles (referenced 'A', 'B' and 'D' circles) illustrate the 21 bread wheat chromosomes based on the BIN locations ('BIN' circle), the genetic map ('cM' circle where 375 mapped COS-SNPs are illustrated as red bars) and 6423 homoeologous genes (green curves). The recombination density heatmap is illustrated as red (for BIN where physical size is higher than cM size) or blue (for BIN where physical size is lower than cM size) BIN intervals. The characterized dominance and sensitivity of the modern wheat genomic compartments are mentioned on the external circle.
(b) Screen capture of the PlantSyntenyViewer web tool [http://urgi.versailles.inra.fr/synteny-wheat] visualizing the synteny between wheat, *Brachypodium*, rice, maize, sorghum and providing the COS marker information (ID, primer pairs, SNP in wheat).

performed two complementary strategies, bypassing the requirement of a complete reference genome sequence, consisting aligning the public wheat gene repertoire to the previous COS-based wheat syntenome (strategy #1 in Figure 1a referenced as 'Wheat→COS') as well as COS *de novo* sequencing in wheat (strategy #2 in Figure 1a referenced as 'COS→Wheat'), Figure S1.

The first strategy (#1, black arrow Figure 1a) in investigating the retention of ancestral genes in bread wheat consisted in comparing the recently published wheat genome shotgun sequence (Brenchley *et al.*, 2012) to the previous wheat syntenome. When comparing the published wheat genome consisting in ~95K gene models with 18 483 of them assigned to the three subgenomes (consisting in 48 120 homoeologous-specific genes distributed 15 996 on A, 15 244 on B and 16 880 on D) to the 17 317 ordered protogenes in wheat (mapped on chromosome groups 1–7), 6423 (35%) ancestral grass genes appear retained in wheat (Table S1 and Figure 1a). This reduced rate of gene conservation is consistent with the 17 093 genes in *A. tauschii* with 26.1% associated with orthologs in grasses (Luo *et al.*, 2013) as well as 42% (14 578 genes) of the 34 879 gene models from *T. urartu* reported as conserved with *Brachypodium* (Ling *et al.*, 2013). Such limited level of single gene orthologous relationships between the wheat gene repertoire and the ancestral retained grass genes in the other species, is largely due to wheat-specific gene content amplification though intra- and inter-chromosomal duplications (Figure S2). The distribution of the retained ancestral genes on the wheat subgenomes (4905, 158, 347, 652, 62, 187, 112 respectively located on A–B–D, A–B, B–D, A–D, B, D and A genomes specifically) showed an increased gene density at the subtelomeric regions compared to the centromeres (compare with BIN-based gene density distribution in green on the Figure 1a). Homoeolog gene density appeared correlated to the recombination rate, as previously reported in the diploid *A. tauschii* (Jia *et al.*, 2013; Luo *et al.*, 2013) and *T. urartu* (Ling *et al.*, 2013) genome sequences. The density of non-colinear genes per bin was correlated with recombination rate, with non-collinear genes located in distal compartments of the chromosomes, then suggesting that non-collinear genes may have survived in distal high recombination regions compared to proximal low-recombination regions.

The wheat syntenome illustrates that the seven Triticeae chromosome groups originated from a $n = 12$ ancestral species by five nested chromosome fusions (NCFs), involving chromosomes 1 (A5 + A10), 2 (A4 + A7), 4 (A3 + A11), 5 (A12 + A9 + A3), 7 (A6 + A8), Figure 1(a). The previous chromosome-to-chromosome relationships, established at the BIN-resolution level between wheat and the sequenced grass genomes, allowed us to transfer the ancestral dominant (A1–2–3–4–9–11, associated with higher retention of ancestral genes) and sensitive (A5–6–7–8–10–12, associated with higher loss of ancestral genes) chromosomal blocks reported in grasses (Figure S3) into the 21 bread wheat chromosomes (Figure 1a with D and S blocks). Among the 157 BINs available on the 21 bread wheat chromosomes, 79 and 69 were respectively identified as dominant and sensitive according to the known nature of their orthologous counterparts in the sequenced grass species (Table S2). The Triticeae genomes, that derived from the $n = 12$ ancestor (centre on the circle on Figure 1a) through five ancestral chromosome fusions, can then be re-organized into paleo-D and paleo-S (*i.e.* paralogous dominant (D) and Sensitive (S) blocks deriving from the ancestral WGD) chromosomal compartments where w1(S) = A10(S) + A5(S), w2(D) = A7 (S) + A4(D), w3(D) = A1(D), w4(D) = A11(D) + A3(D), w5 (D) = A9(D) + A12(S), w6(D) = A2(D), w7(S) = A8(S) + A6(S).

Such high resolution BIN-based wheat syntenome offered the opportunity to investigate the retention of the ancestral genes in wheat following the paleotetraploidization event (by comparing D and S wheat blocks defined previously) as well as the neohexaploidization event (by comparing the A, B and D subgenomes), as illustrated in Figure 2(a) (top) with D chromosomes in blue and S chromosomes in red. Regarding the number of retained genes in D vs. S blocks, we observed the expected higher retention ($P < 5\%$) of protogenes in paleodominant blocks compared to paleosensitive counterparts (Figure 2b, left). This result confirm the existence of shared D and S blocks in wheat following the paleoteraploidization (~65 mya) as reported in rice, sorghum, *Brachypodium* and maize (Abrouk *et al.*, 2012). Surprisingly, when comparing the



retention of protogenes in the A, B and D subgenomes, we also observed a bias retention ($P < 5\%$, Figure 2b right) of ancestral genes, with the B subgenome showing the property of a so-called sensitive subgenome with higher loss of protogenes. This observation raised the hypothesis that the reported genome partitioning phenomenon following the ancestral shared paleotetraploidization event (bias retention of protogenes between paleo-D and S blocks confirmed in wheat as illustrated in Figure 2b left) may also acts between the A, B and D subgenomes deriving form the neohexaploidization, with A and D as dominant (i.e. stable) and B as sensitive (i.e. plastic), Figure 2(b) right. This confirms and largely complements recent evidences of differential gene loss observed between the subgenomes of the wheat chromosome group 7 with increased gene numbers reported respectively on D > A > B (Berkman *et al.*, 2013).

However, the previous conclusions regarding the wheat genome partitioning in response to both paleo- and neopolyploidization events were obtained in exploiting the public wheat genome shotgun sequence (Brenchley *et al.*, 2012) that consists only in a partial gene repertoire. We cannot exclude that missing (non-sequenced) genes/families/homoeologs in such resource may have impacted our conclusion regarding the observed structural plasticity of paleo- and neoduplicated compartments in modern hexaploid bread wheat. Consequently, in order to validate and complement the previous conclusions regarding the wheat subgenome dominance, we performed a second strategy (#2, black arrow Figure 1a and Figure S1) that consisted in *de novo* sequencing and mapping of COS genes in wheat deriving from the 17 317 protogenes but not included in the previous 6423 wheat/grass ortholog repertoire (Figure S4). To do so, we applied a COS-finder tool (see Method, Quraishi *et al.*, 2009) to the 10 894 (17 317 total COS excluding the 6423 COS identified from Brenchley *et al.*, 2012) ancestral grass genes not associated with a public wheat gene model sequence. We have then been able to develop a tremendous catalog of 6033 primer pair set (primers pairs selection criteria in Method section) defining 5234 COS relationships (Table S1). Such COS primers, selected on highly conserved exonic regions, offer all the guarantee to be useful in sequencing the wheat orthologs (and associated homoeologs), based on the observed transferability of such COS markers among monocot species (i.e. sorghum, maize, oat, rice, Triticale, wheat, *Brachypodium* and rye) at both the amplification (Figure S5a on agarose gels and Figure S5b on capillary sequencer) and sequence levels (Figure S5c with melting curve profiles supported by sequence-based haplotyping data). Overall, the characterized COS markers (17 317 COS with 5234 selected COS for sequencing and 6033 associated primer pairs) as well as the COS-finder software are made available at http://urgi.versailles.inra.fr/synteny-wheat (Figure 1b and Table S1).

In order to validate the orthologous, homoeologous and paralogous gene contents and retention in wheat paleohistory, raising the evolutionary model suggesting subgenome dominance between D>A>B, we performed the sequencing of the selected 5234 COS (Experimental Procedure and detailed strategy in Figure S6). COS markers have been sequenced in *Chinese Spring* using the 454 (Roche) technology, delivering 23 463 reference transcripts (RTs) covering 4582 COS (with an average of three RTs per COS, putatively homoeoalleles, Figure S6). This result established that 88% of the COS (i.e. genes conserved in rice, *Brachypodium*, maize and sorghum) are conserved in wheat, in contrast to 35% of the 18 483 gene models (from Brenchley *et al.*, 2012) that have been associated with a COS gene in the previous section. This result reinforces the previous conclusion (Figure S2) that wheat has been enriched, in the course of evolution, by species and/or lineage-specific genes. The capture (Agilent SureSelect Target Enrichment) of the 23 463 RTs in eight bread wheat genotypes (*Chinese Spring*, *Renan*, *Courtot*, *Alcedo*, *Brigadier*, *Genial*, *Hustler* and *Nicam*) and short read sequencing (Illumina HiSeq 2000) delivered 9969 high quality SNP covering 2197 COS markers with a density of 1.9 SNP/500 bp (Table S1 and Figure S7). Such *de novo* COS sequences in wheat allowed us to confirm the bias retention of protogenes defining dominant as well as sensitive blocks for both paleo- and neopolyploidization events (Figure 2c). The number of sequenced COS in wheat observed in dominant and sensitive compartments confirmed our previous conclusions based on the public gene repertoire, where more retained ancestral genes (or orthologs) are observed in the dominant blocks (Figure 2c left). We also confirmed the bias retention of homoeologs between A, B and D subgenomes (Figures 2c right and S8). Overall, we then proposed, based on a pure *in silico* experiment (alignment-based synteny analysis) complement by a *de novo* sequencing strategy (COS characterization), that the ancient subgenome dominance process related to the shared paleotetraploidization in grasses is observed in wheat and has been eroded (*P*-values from $\sim 3.10^{e-2}$ to $\sim 5.10^{e-2}$) by the neohexaploidization event that may have led to a modern subgenome dominance where B is sensitive and A–D are dominant (*P*-values from $\sim 1.10^{e-3}$ to $\sim 8.10^{e-7}$).

Genome partitioning shaped wheat genetic landscape

In order to address not only the bias in gene retention (detailed in the previous section) but also in gene diversity between modern and ancient dominant *vs.* sensitive blocks, we used the overall depth of read coverage across COS as well as the depth of read coverage at variable sites to detect not only sequence variants (i.e. 9969 SNPs, Figure S7) but also putative structural variants such as presence/absence variations (PAVs), copy number variations (CNVs)

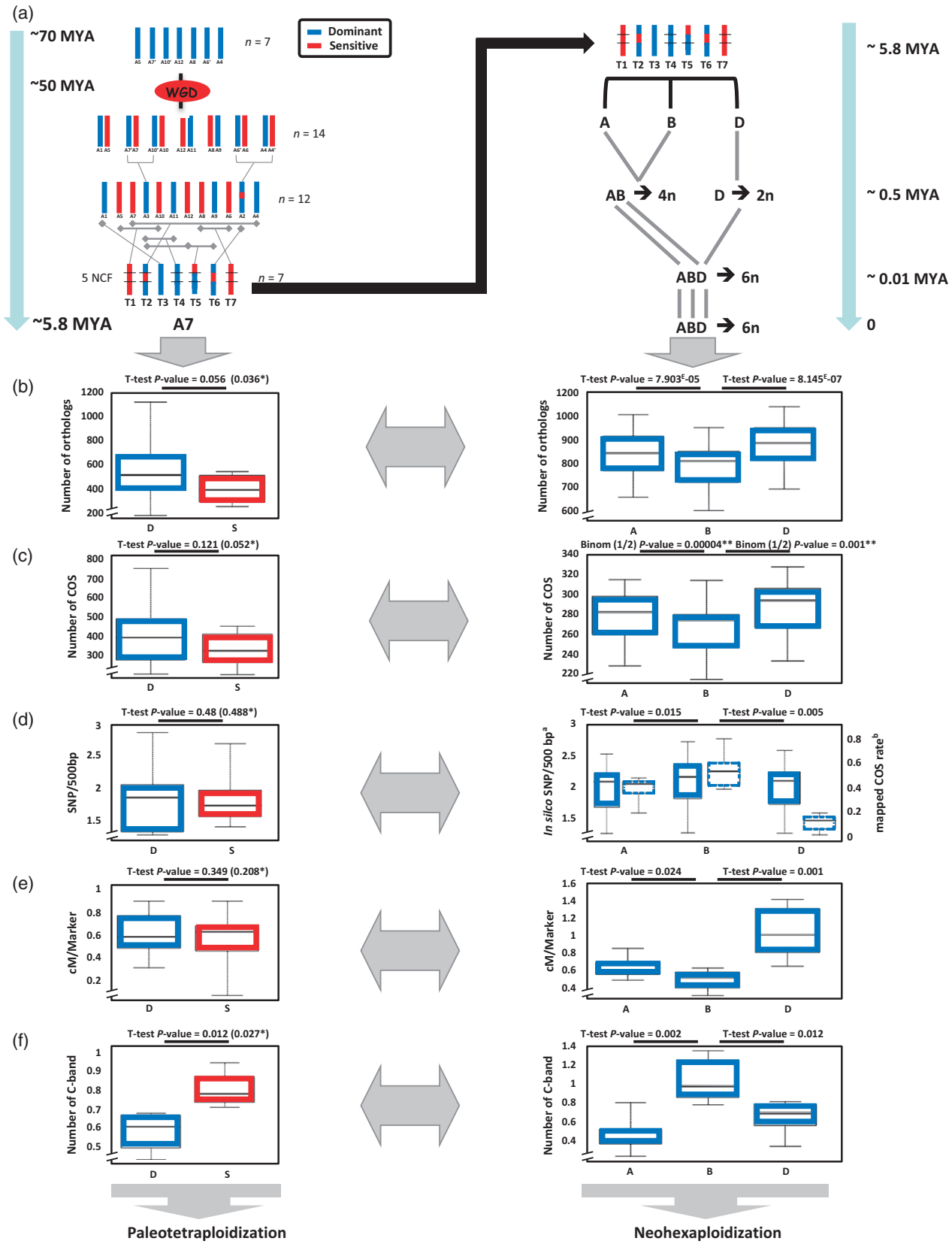
Table 1 Wheat syntenome and derived COS-SNP makers

	Gene repertoire				COS genes			COS-SNP markers		
Chr.	Genes/ group	gA	gB	gD	Protogenes	COS primers	COS-seq[a]	COS-SNP[b]	COS Mapped[c]	COS conserved	COS transposed[c]
1	1945	655	600	690	2001	698	634 (2792/42/64)	1492 (299)	76 (36/28/12)	72 (94%)	4 (0/4/0)
2	2964	882	942	1028	2963	1020	882 (4139/41/125)	2077 (386)	64 (27/33/4)	57 (89%)	7 (6/1/0)
3	2830	755	880	993	2733	924	755 (3298/40/105)	1507 (323)	82 (25/47/10)	77 (93%)	5 (1/3/1)
4	2553	770	809	889	3019	776	685 (3035/39/79)	861 (243)	18 (7/11/0)	12 (66%)	6 (4/2/0)
5	2519	788	804	878	2023	941	873 (4012/59/103)	1339 (352)	28 (5/22/1)	15 (53%)	13 (1/12/0)
6	2156	625	681	763	2097	717	625 (2769/42/69)	1147 (264)	51 (22/22/7)	49 (96%)	2 (0/1/1)
7	2423	780	756	850	2481	957	780 (3418/41/99)	1546 (330)	56 (25/21/10)	53 (94%)	3 (2/1/0)
Total	17 390 (48 120)[d]	5827 (15 996)[d]	5472 (15 244)[d]	6091 (16 880)[d]	17 317	6033	5234 (4286/304/644)	9969 (2197)	375 (147/184/44)	335 (89%)	40 (14/24/2)

[a]In parenthesis: number of NSV, PAV and CNV.
[b]In parenthesis: number of COS corresponding to the number of SNP.
[c]In parenthesis: number of mapped and transposed COS respectively in A, B and D subgenomes.
[d]In parenthesis: total number of wheat homoeologous genes from Brenchley *et al.* (2012).

in contrast to non-structural variation (NSV), in which all genotypes are covered for the considered COS). Figure S9 illustrates three distinct COS markers associated with: (i) sequences for the eight sequenced genotypes [COS-5368, observed in 80% (4286 COS) of the cases]; (ii) missing sequences [COS-542, PAV observed in 6% (304 COS) of the cases]; or (iii) extra copy sequences [COS-3070, CNV observed in 14% (644 COS) of the cases]. We then provide a complete view of the gene diversity distribution in wheat based on the precise characterization of both non-structural (SNPs) as well as structural variants (InDels, CNVs and PAVs). The wheat COS sequences offered the opportunity to test the impact of the reported biased chromosomal plasticity between dominant and sensitive blocks on gene diversity through differential SNP/InDel contents observed in such wheat genomic compartments. Figure S10 illustrates the observed distribution of InDel and SNP polymorphisms in wheat where the more frequently observed (61% of the cases) sequence variations of SNPs correspond to transitions (A/G or T/C) compared with transversions (A/T, A/C, T/G or G/C). The most frequently observed (52% of the cases) sequence variations of InDels correspond to a single nucleotide. The distribution of SNPs on D and S blocks as well as A, B and D subgenomes shows that the gene diversity increases in the sensitive blocks (either paleo-S blocks or B genome, Figure 2d, plain box-plots). Wheat chromosome groups derived five fusions of the 12 ancestral chromosomes then juxtaposing high-gene-density terminal paleoregions near low-gene-density paleocentomeric regions in the post-fusion neo-chromosome (Figure 1a doted connecting black lines), Murat *et al.* (2010). Consequently, despite bias distribution of sequence variants between dominant and sensitive subgenomes, we observed a higher gene (COS) diversity (i.e. SNP/bp density) in telomeric regions and synteny breakpoints (SBP) regarding the wheat chromosome groups that derived from the centromeric nested insertion of ancestral chromosomes (Figures 1a black stars and S11).

Until the wheat genome is entirely sequenced and available (in delivering complete pseudomolecules) the discovery and application of genome polymorphism for wheat crop improvement and evolutionary investigation remains a real challenge, in which the provided wheat syntenome constitutes a tremendous matrix for these purposes. Based on the previous *in silico* syntenome and SNP-based diversity map, we addressed differences in gene content and plasticity between dominant and sensitive fragments for both paleo- and neopolyploidization events. As the previous conclusions derived from pure *in silico* syntenome and COS-SNPs characterization, we validated the delivered computed wheat gene order (17 317 ordered COS) as well as associated sequence variants (9969 SNPs) through COS-SNP mapping (Figure S4). As a validation procedure, we have randomly selected a subset of 1135 (22%) among the



5234 COS markers to test the accuracy of the characterized *in silico* COS-SNPs as well as computed gene order of such COS. We have identified 807 *in silico* SNPs from 986 (out of 1135, 87%) COS markers that have been successfully sequenced in genotypes for witch mapping populations are available (see Experimental Procedure section). Five hundred and forty high quality scoring (Quality score A to D) and 267 low quality scoring (Quality score classes E and F) *in silico* SNPs (based on SNP calling criterion described in Figure S7) have been tested. 357 (66%) and 63 (24%) respectively have been validated as polymorphic between bread wheat cultivars *Chinese Spring*, *Courtot*, *Arche*, *Recital* and *Renan*. 375 (89%) out of the 420 (357 high quality + 63 low quality scores) validated SNP have been successfully mapped (Illumina approach) on the three available mapping populations (*Courtot* × *Chinese Spring*, *Arche* × *Recital*, *Renan* × *Recital*, see Experimental Procedure), illustrated in Figure 1(a) (red bars on the wheat chromosome circles) and available in Table S1. Such observed validation rate (420/807 = 52%) of *in silico* SNPs is consistent with the one (60%) reported by Allen *et al.* (2011) as well as in Trick *et al.* (2012) (56–58% of validation accuracy). Overall, based on the previous genotyping data, we provided in the current analysis 9969 high confidence SNPs (deriving from 5234 sequenced COS markers) for witch 4613 (9969 × 0.52 × 0.89) are expected to be converted efficiently in any functional assays. The remaining non-polymorphic *in silico* SNPs may be related to: (i) heterogeneity observed in using different genotyping approach (for example five SNPs failed to be mapped using Illumina technology but recovered with KASpar); (ii) *in silico* intervarietal high quality SNPs that happened to be homoeoSNPs; and (iii) SNPs associated with the same quality score that can either be validated or not through Illumina genotyping approach (Figure S12).

COS-SNPs have been mapped and tested on Illumina, KASpar, SSCP, HRM, LNA detection (see Experimental Procedure) to test the accuracy of the provided computed wheat gene order based on the observed COS genetic position, Figure 2(b). Overall, taking into account the mapped COS-SNP information over the 21 bread wheat chromosomes, 89% (335) of mapped COS were retained at the conserved (i.e. expected orthologous position) loci whereas 40 are non-syntenic (Table 1 and Figure S13). Whereas 147 and 184 COS-SNPs have been mapped on the A and B genomes, only 44 have been positioned on the D-genome (Table 1 and illustrated as 'mapped COS rate' on Figure 2d, dashed box-plots) supporting the higher plasticity of the B (sensitive) compared to the A (dominant) related to long-term paleo-tetraploidization evolutionary event; and between A/B and D subgenomes due to the low level of genetic diversity in the D-genome associated with recent origin of hexaploid wheat. Comparison of the *in silico* COS-SNP density (Figure 2d, plain box-plots representing the observed 'SNP/500COS ratio' from 9969 *in silico* COS-SNPs) and mapped COS-SNP rate (observed from the 375 mapped COS-SNPs) it appeared that the D genome diversity may has been overestimated *in silico* through the possible mis-association of SNPs either on the A or D genomes, especially for short sequence reads not harbouring clear and multiple homoeoSNPs (Quality score classes E and F, Figure S7). Interestingly the 11% (40 genes) of COS-SNP markers mapped at non-orthologous positions (that can be considered as transposed genes) are mostly located (60%) on the B genome, reinforcing the structural plasticity nature of such genomic compartment (Table 1). As expected, due to the considered time frame (~65 mya), bias in gene diversity (expressed as SNPs/bp) cannot be observed for the ancestral paleotetraploidization event (Figure 2d left). The same observation is made when investigating the impact of paleo-dominance and paleo-sensitivity on the recombination pattern where no difference can be observed between paleo-D vs. paleo-S blocks ($P > 5\%$) but a higher cM/marker ratio was observed for the B compared with the A ($P = 2.4 \times 10^{-2}$) and the D ($P = 1 \times 10^{-3}$) subgenomes (Figure 2e). The observed reduced cM/marker ratio of D subgenome compared with the A and B counterparts has been attributed to the loss of polymorphism during the genetic bottleneck that accompanied the development of modern elite cultivars. Overall, taking into account the provided COS-SNP markers associated with reference public wheat genetic maps, we then produced the most complete composite bread wheat genetic map (consisting in 7520 molecular markers covering 4318.03 cM with a marker density of one marker every 0.78 cM) of the 21 chromosomes and including mapped RFLP (Restriction fragment length polymorphism) (1687), AFLP (712), STS (262), SSR (2315), DaRTs (1246) and COS-SNP (375) gene markers, named Wheat Composite Genetic Map

Figure 2. Wheat paleo- and neosubgenomes.
(a) Illustration of the dominant (blue) and sensitive ancestral chromosome derived from a $n = 7$ AGK that has been duplicated (WGD) to reach a $n = 12$ intermediate. The Triticeae ancestor (T1–T7) derived from five NCF that took place between the 12 ancestral chromosomes (left). The modern hexaploid bread wheat derived from the polyploidization between three diploid progenitors (A, B and D sub-genomes) structured in seven protochromosomes (right). Distribution and associated statistical significance are illustrated as box-plots (considering individually the seven wheat chromosomal groups) for the number of orthologs (b), number of COS (c), SNP (aSNP/500COS ratio from 9969 *in silico* COS-SNPs and bmapped COS rate from 375 mapped COS-SNPs) data (d), cM/marker ratio (e), number of C-band (f) between ancestral D and S blocks (left) as well as modern A, B and D subgenomes (right) in wheat. Paired *t*-test *P*-values are referenced between paleo-D and paleo-S blocks (at the left, with A11–A12 pair excluded in parenthesis highlighted with *), and between A–B and B–D subgenomes (at the right, with binomial ($n = 2$) *P*-value highlighted with **).

Figure 3. Wheat evolutionary model.
(a) Evolutionary model of the modern wheat genome from a $n = 7$ (AGK), 12 (duplicated AGK), 7 (Triticeae) ancestors illustrated as dominant (blue bars) and sensitive (red bars) blocks. The subgenome dominance took place in wheat paleohistory when the diploid progenitors hybridized. Genome A (blue outlines) was dominant and B (red outlines) sensitive after the first polyploidization event 1.5 mya. The tetraploid (A–B genome) was sensitive and D dominant after the second polyploidization event 0.1 mya. The modern 21 bread wheat chromosomes are then illustrated as dominant, sensitive, supra-dominant and supra-sensitive fragments according to the colored scale at the right defining distinct degree of genomic stability and plasticity.
(b) Distribution of gene number, ortholog number, chromosome length, transposable element observed between paleo-S and paleo-D fragments in rice (left). Distribution of COS number, SNP/COS ratio, cM/marker ratio, number of transposed genes, C-band number observed between chromosomes 1A (sensitive), 1B (supra-sensitive), 1D (sensitive), 3A (dominant), 3B (dominant), 3D (supra-dominant), right.

'WCGM2013' (available as Table S3 and detailed provided in Figure S14).

Genome partitioning in delivering D and S fragments (shown as associated to bias in gene content, diversity and recombination in the current study) following polyploidy has been suggested to be epigenetically driven (Doyle *et al.*, 2008). To test this hypothesis, we investigated C-band (from Hutchinson *et al.*, 1982) differences between paleo- and neodominant and sensitive blocks (Figure 2f). Where differences are observed between ancient D and S compartments $P = 2.7 \times 10^{-2}$, Figure 2f left), the difference is more visible between the neodominant (A–D subgenomes) and neosensitive (B subgenome, $P = 1.2 \times 10^{-2}$ and 2×10^{-3} respectively) compartments, Figure 2(f) right. The transition nature (A/G or T/C) of the reported 9969 SNPs may also indicate the historic methylation profile of each subgenomes as the transition abundance bias is commonly observed for SNPs reflecting the high frequency of C to T mutation following methylation (Coulondre *et al.*, 1978). Such a bias observed in polyploid wheat is greater than that observed in its diploid Triticea relative barley (Duran *et al.*, 2009) and may reflect a higher level of methylation in polyploid wheat genome as part of the diploidization process leading to D (low methylation) and S (high methylation) compartments. Both observations (C-band differences as SNP transition enrichments) lead to the opening question of epigenetic driving the observed paleo- and neodominance in wheat as well as in other grasses. Overall, the clear characterization of ancient and recent D-S compartments, in relation respectively to the paleotetraploidization and the neohexaploidization events, suggest that they may have shape the modern bread wheat genomic and genetic landscapes with bias in gene number (i.e. higher gene retention in dominant blocks), gene diversity (i.e. higher variant per genes in sensitive blocks), recombination (i.e. lower cM/Marker in sensitive blocks) and heterochromatin structure (i.e. higher C-band and possibly methylation in sensitive blocks).

Bread wheat evolutionary model

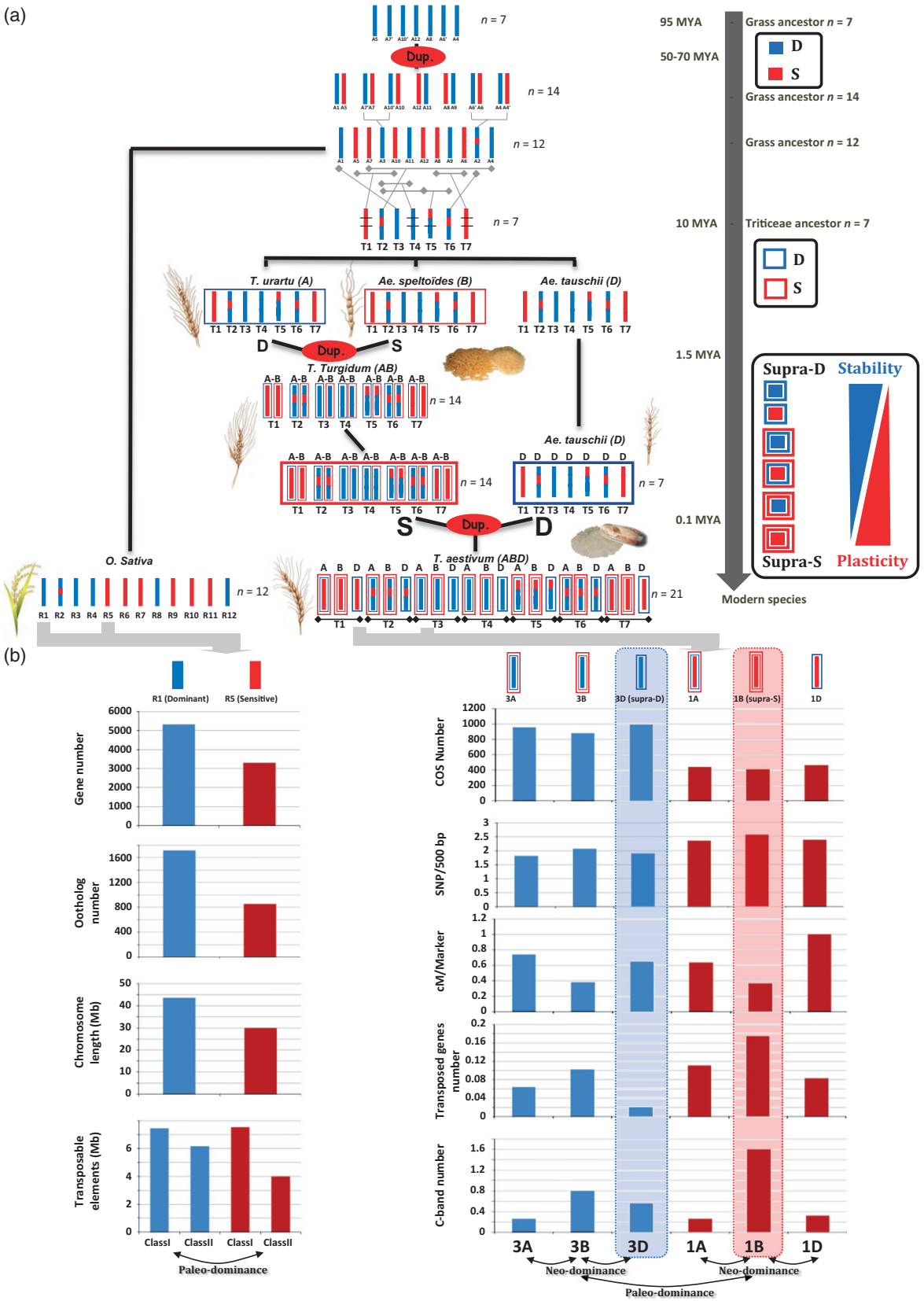
Based on the previous observations, we propose an evolutionary scenario that has shaped the modern bread wheat genome during the last 65 my of evolution in response to three polyploidization events (Figure 3a). The $n = 7$ AGK has been paleotraploidized leading to seven pairs of dominant (red) and sensitive (blue) chromosomes. Five ancestral chromosomes fusions took place to build the seven Triticeae chromosomes and wheat ancestors (i.e. *T. urartu*, *A. speltoides* and *A. tauschii*), where T3–T4 are entirely sensitive, T1–T7 are entirely dominant and T2–T5–T6 are structured as a mosaic of D and S blocks. The first neotetraploidization event (1.5 mya) leaded to a subgenome dominance where the A subgenome became dominant and the B subgenome sensitive. The second neohexaploidization event (0.1 mya) leaded to a supra-dominance where the tetraploid became sensitive (subgenomes A and B) and the D subgenome dominant. Thus, the modern bread wheat genome can be divided into five supra-dominant chromosomes (also referenced as pivotal) deriving from surimposed dominances (i.e. chromosomes 3D, 2D-S/L, 4D, 5D-L, 6D-S/L) in contrast to five supra-sensitive ones (i.e. chromosomes 1B, 2B-C, 5B-S, 6B-C, 7B), and the remaining regions are associated with opposite D and S following the successive duplication events (i.e. with S for short arm, L for long arm and C for centromeric region), see Figure 3(a) (with the stability/plasticity scale at the right).

In order to investigate precisely supra-dominant and supra-sensitive chromosomal architecture, we focussed on rice chromosomes 1 (dominant) and 5 (sensitive) associated with differences in gene content, orthologous/ancestral gene retention, chromosome length as well as transposable element (mainly TE class II) repertoire (Figure 3b left). The orthologous counterparts in wheat (group 1 as sensitive and group 3 as dominant) display the same bias in number of orthologs (i.e. COS number higher on the dominant group 3) as well as gene diversity (i.e. SNP/COS density higher on the sensitive group 1). In our scenario, among the six investigated chromosomes, 3D is supra-D in contrast to 1B as supra-S. Consequently the chromosome 1B displays the lowest gene retention, the highest gene diversity, lowest cM/marker ratio, the highest number of transposed genes and highest number of the C-band.

DISCUSSION

Wheat genome architecture has been shaped by subgenome partitioning following paleo- and neopolyploidization

Common anchors (COS genes) to compare genomes and reconstruct CARs are necessary to unravel plant genome paleohistory (Salse, 2012). It has been shown that CARs duplication in the grass paleohistory has been followed by



a structural partitioning in defining post-duplication dominant regions (defined as structurally stable with higher retention of protogenes) in contrast to sensitive paralogous counterparts (defined as structurally plastic with higher loss of protogenes), Abrouk *et al.*, 2012; Schnable *et al.*, 2012a,b. Based on the chromosome-to-chromosome synteny relationships established between the seven bread wheat chromosome groups and the rice, sorghum, *Brachypodium* and maize genomes, it was then possible to produce a partial wheat gene-based physical map (i.e. syntenome) including 17 317 COS. However, in the absence of large-scale gene mapping data in wheat, how relevant such computed gene order is robust, remains an open question and is of importance if one wants to investigate the evolution of wheat gene content in comparison of the sequenced grasses genomes. Our current study aiming at sequencing and mapping ancestral COS genes in bread wheat established that the synteny-based computed gene order delivered is currently correct for up to 89% of the simulated gene order and that lineage-specific rearrangement in wheat may account for <11% of the ordered genes *in silico*.

Such wheat syntenome, that deliver *in silico* synteny-based ordered genes conserved in close relatives, is a perfect matrix to characterized genes that have been conserved or lost in bread wheat evolution in response to paleotetraploidization (~65 mya) and neohexaploidization (>1.5 mya) events. Using complementary strategies consisting in aligning the public wheat genome repertoire (Brenchley *et al.*, 2012) as well by *de novo* COS gene sequencing in wheat, we identified respectively 35% of wheat genes associated with an ortholog in the other grasses and 88% of COS successfully sequenced in wheat, suggesting that lineage-specific gene amplification and shuffling mechanisms have shaped the wheat genome in its recent evolution. More interestingly, more conserved genes are observed in dominant wheat regions in contrast to sensitive ones arising from the ancestral shared paleotetraploidization event as reported in the other grass species (Abrouk *et al.*, 2012). Moreover, such bias in gene conservation is also observed between the three wheat subgenomes deriving from the neohexaploidization, with more orthologs detected in D > A > B. This subgenome dominance phenomenon may then explain the observed differences in genetic and physical size reported between subgenomes, the B (plastic) as the largest and the D (stable) the smallest (Furuta *et al.*, 1986). Bias retention of ancestral genes has been observed between paleoduplicated (plasticity of S > D) as well between neoduplicated (plasticity of subgenomes B > A > D) genes, suggesting that ancient as well as recent polyploidization events are followed by a diploidization mechanism consisting in the structural partitioning of the paralogous fragments.

Despite the characterization of conserved genes in the wheat subgenomes leading to the identification of paleo- as neodominant and sensitive chromosomal regions, we investigated non-syntenic genes in wheat corresponding: (i) at the genome level to lineage-specific intra- (18%) and inter-chromosomal (82%) duplications; and (ii) at the gene level to CNVs (14%), PAVs (6%) as well as gene transposition (20%). As the COS markers have been selected has single copy conserved genes, it is unlikely that all cases of non-syntenic genes were due to erroneous mapping of duplicates or homologs. That raises the possibility of a small scale (i.e. few genes) transposition/movement mechanism despite large-scale ones (such as translocations and inversions) involving DNA segments carrying several genes, operating in the decay of synteny between grass genomes and then driving non-conserved genes in wheat (i.e. ~50% initially reported in wheat from Pont *et al.*, 2011; and ~60% in the current analysis). More transposed genes have been observed on the B subgenome (sensitive according to the current analysis), that is consistent with the conclusion raised on chromosome group 1 (Wicker *et al.*, 2011), based on partial sequence investigations. Gene movement may appear as a particularly active phenomenon in polyploidy wheat and may act then preferentially on the sensitive compartments.

Revisiting the B genome progenitor enigma based on contrasted plasticity following polyploidization

Several wheat phylogeny studies have tried to identify the progenitor of the B genome of polyploid wheat based on cytology (Zohary and Feldman, 1962), nuclear and mitochondrial DNA sequences (Dvorak *et al.*, 1989; Dvorak and Zhang, 1990; Terachi *et al.*, 1990) as well as chromosome rearrangement studies such as common translocation events (Feldman, 1966a,b; Hutchinson *et al.*, 1982; Gill and Chen, 1987; Naranjo *et al.*, 1987; Naranjo, 1990; Jiang and Gill, 1994; Devos *et al.*, 1995; Maestra and Naranjo, 1999). More recent and representative molecular comparisons, at the whole-genome level using germplasm collections, have shown that the B genome could be related to several *A. speltoides* lines but not to other species of the *Sitopsis* section (Salina *et al.*, 2006; Kilian *et al.*, 2007). At the *SPA* locus level (Salse *et al.*, 2008), close relationships between the *A. speltoides* and the hexaploid B subgenome has been reported based on both coding and non-coding sequence comparisons, but with a lower conservation compared to the A subgenome and its *T. urartu* progenitor at the PSR920 region (Dvorak and Akhunov, 2005; Dvorak *et al.*, 2006). The greater genetic diversity of the B genome compared to the A genome of hexaploid wheat was also reported based of SNPs (previous references), SSRs (Roder *et al.*, 1998a,b), RFLPs (Liu and Tsunewaki, 1991), SNPs (Akhunov *et al.*, 2010), as well as in tetraploid (Thuillet *et al.*, 2005; Ren *et al.*, 2013). Such contrasted conservation between B subgenome and *A. speltoides* compared with

the A genome and its *T. urartu* progenitor is classically explained with two hypothesis where: (i) the progenitor of the B genome is a unique *Aegilops* species that remains unknown (i.e. monophyletic origin and ancestor closely related to *A. speltoides* from the *Sitopsis* section); or (ii) this genome resulted from an introgression of several parental *Aegilops* species (i.e. polyphyletic origin) that need to be identified from the *Sitopsis* section. Overall the B genome is therefore with greater divergence compared to the A and D subgenomes relative to their putative diploid progenitors, with the D subgenome as the most stable in the course of evolution. The contrasted genomic (gene loss and diversity) and genetic (recombination and Linkage Disequilibrium) plasticity reported for the B > A > D (from the most plastic to more stable subgenomes) have been tentatively explained so far through hypotheses relying on diploid progenitors differences. However, these hypotheses rely on the assumption of a constant and similar evolutionary rate between subgenomes.

In contrast, we propose an alternative scenario where such particular conservation of the B subgenome with *A. speltoides* at the sequence level is the consequence of a differential evolutionary plasticity of the B subgenome compared with the other A and D subgenomes in response to polyploidization events. We then propose an evolutionary scenario where the modern bread wheat genome has been shaped through a first neotetraploidization event (1.5 mya) leading to a subgenome dominance where the A subgenome was dominant and the B subgenome sensitive. The second neohexaploidization event (0.1 mya) leaded to a supra-dominance where the tetraploid became sensitive (subgenomes A and B) and the D subgenome dominant (i.e. pivotal). Following our scenario, wheat subgenome architecture has then been polyploidy-driven (i.e. pure post-polyploidization mechanism), where genome doubling is followed by genome portioning into dominant (stable) and sensitive (plastic) compartments, instead of entirely explained by pre-existing differences between founder (A, B and D) progenitors (i.e. pure pre-polyploidization mechanism). We propose in the current study that subgenome dominance is not only active in paleopolyploids but also in modern neopolyploids such as bread wheat. In our bread wheat evolutionary model, we propose that the differences in genomic structure and genetical landscape between subgenomes resulted directly from differential modes of evolution between dominant and sensitive blocks, where B subgenome acts as a sensitive in contrast with A as dominant in response to the neotetraploidization event 2.5 mya and D as dominant and A–D subgenomes as sensitive in regard to the neohexaploidization event 10 000 years ago.

This scenario is in agreement with the differential plasticity (at the gene content and diversity levels) reported in the current analysis in regards to gene loss and SNP/gene, higher in the B subgenome compared to the two others. Moreover, our wheat subgenome dominance model provides highlight into earlier studies of genome rearrangement (Zhang *et al.*, 2013) or gene loss (Ozkan and Feldman, 2001b; Ozkan *et al.*, 2001a) in newly synthesized polyploids as well as the reported biased erosion of genetic diversity during domestication (Cavanagh *et al.*, 2013). The B genome, sensitive subgenome in regard to the paleotetraploidization, is associated to a pronounced genome plasticity related to the number of paralogous loci (reported higher than in the A and D genomes in Akhunov *et al.*, 2003), translocation (Kota and Dvorak, 1988), and other large-scale structural changes (Zhang *et al.*, 2013). Zhang *et al.* (2013) using a set of 16 independently synthesized allohexaploid wheat lines produced >1000 individual plants at different selfing generations (S1 > S20) after allohexaploidization that have been karyotyped. Of the three consistent subgenomes, B (sensitive) showed the highest frequency of chromosome rearrangements, followed by A (dominant); and the D (supra-dominant, referenced also as 'pivotal' in Zhang *et al.*, 2013) subgenome was largely observed with the higher stability. Overall, these data suggest that DNA rearrangements at the chromosome as well as gene levels, leading to the reported diploidization-driven subgenome dominance, occurred immediately or within a few generations following polyploidization.

Both scenarios aiming at explaining the structural asymmetry characterized between the wheat subgenomes need to be now considered, i.e. either a pre-polyploidization mechanism where the evolutionary history of the cross-pollinator *A. speltoides* progenitor (B donor) had a more diverse genome that self-pollinators *T. urartu/A. tauschii* (A/D donors) or a post-polyploidization mechanism with an accelerated plasticity of the B subgenome in contrast to the homoeologous counterparts.

Syntenome of complex polyploid species can be used as a guide for trait dissection

High resolution and large-scale comparative genomics studies offer a tremendous set of gene-based markers that can be used directly as founder resource for genome mapping (physical or genetic) and ultimately trait dissection. In the present work, we deliver a wheat syntenome consisting in 17 317 ordered COS genes in wheat and the associated sequence variations (SNP and InDels). Overall, the sequence capture of the COS markers associated with NGS methods appear to be a powerful technique for the large-scale discovery of gene-associated SNPs in any plant of interest. Moreover, such sequencing efforts within a large-scale intra-specific background will provide a complete set of putative causal SNPs, PAVs, CNVs, as functional diagnostic markers in the near future. Associated with public RFLP, AFLP, STS, SSR, DaRTs markers, COS-SNP characterization and mapping effort in wheat allowed us to provide

here the most complete genetic map consisting in 7520 molecular markers. Such resolution including synteny-based COS markers allows to refine metaQTL intervals and immediately benefits from the COS markers to access robust links with sequence genomes and precise orthologous candidate genes as we precisely illustrated in Quraishi *et al.* (2011a,b) and Dibari *et al.* (2012). Moreover the wheat syntenome has been also used as a matrix for physical map construction (Lucas *et al.*, 2013) prior QTL dissection.

The proposed impact of polyploidization-based subgenome partitioning on contrasted gene content and diversity in dominant and sensitive blocks may need to reconsider in agronomic traits dissection. First of all, the sensitive chromosomal compartments that appeared most plastic (in tem of polymorphism as well as recombination pattern) may be more accessible to QTL cloning. It would be then interesting to investigate the differential efficiency in trait or gene cloning observed in both D and S compartments. Moreover, the difference in subgenome stability (with higher plasticity observed for B > A > D) may also lead to the hypothesis that homoeologous groups may support different type of agronomic traits. Feldman *et al.* (2012a), Feldman and Levy (2012b) reported that genome A was found to control morphological traits while genome B in allotetraploid to control reactions to biotic and abiotic factors. This difference in trait natures in also consistent with the subgenome dominance hypothesis where the sensitive compartment (B subgenome in wheat as defined in the current analysis) has been suggested to control adaptative traits in contrast to more stable processes driven by the dominant counterpart (A and D in hexaploid wheat), Abrouk *et al.*, 2012;. Moreover, QTL partitioning following polyploidy has been also suggested recently with 21% homoeologous fiber quality QTLs in cotton (Rong *et al.*, 2007) and 23% homoeologous fruit quality QTLs in strawberry (Lerceteau-Köhler *et al.*, 2012) characterized, then suggesting that the majority of QTL are no longer maintained on the duplicated blocks as putatively a direct consequence of the diploidization mechanism. To what extent the hexaploid wheat adaptation (in particular regarding adaptation in responses to biotic and abiotic stresses) is possibly partitioned in the genome, between the currently defined dominant and sensitive chromosomal compartments, remains an open question that still needs to be addressed in the future.

EXPERIMENTAL PROCEDURE

The experimental procedures may be found in the online version of this article in detailing: (i) *Wheat Syntenome Protocol* (COS identification; CAR identification; computational gene order in wheat; D-S blocks identification); (ii) *Wheat COS Sequencing Protocol* (COS-primer design; COS-SNP sequencing and clustering); (iii) *SNP Genotyping Protocol* (SSCP, Illumina, size polymorphism, sequencing, KASpar, LNA® Dual-Labeled Fluorogenic Probes); and (vi) *Wheat Comprehensive Consensus Genetic Map Protocol* (selection of public reference genetic maps; Construction of WCGM2013).

FUNDING AND ACKNOWLEDGEMENT

This work has been supported by grants from INRA ('Génétique et Amélioration des Plantes' reference: 'Appel d'Offre AIP Bioressources'), the Auvergne region ('Pôle de compétitivité: Céréales Vallée' reference: programme 'Semences de Demain'), the Agence Nationale de la Recherche (Programs ANRjc-PaleoCereal, ref: ANR-09-JCJC-0058–01 and programme ANR Blanc-PAGE, reference: ANR-2011-BSV6–00801), and the European 7th Framework Programme (programme 'TRITICEAE GENOME', reference: FP7–212019). The author would like to thank Emmanuelle Lagendijk, Nils Stein, Laura Rossini for the coordination of the COS sequencing and Grain Fiber Content QTL characterization initiatives as well as Sébastien Faure for providing the wheat lines in the frame of the TRITICEAE GENOME project.

SUPPORTING INFORMATION

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

Table S1. Wheat syntenome raw data.

Table S2. List of dominant (D) and sensitive (S) chromosomal blocks in wheat.

Table S3. Wheat Composite Genetic Map (WCGM 2013).

Figure S1. Wheat syntenome characterization flow chart.

Figure S2. Wheat duplicated genes (homoeologs & paralogs) characterization.

Figure S3. Grass dominant and sensitive compartments following ancestral WGD.

Figure S4. Strategy for COS-SNP marker development and exploitation.

Figure S5. COS-SNP transferability in sorghum, maize, oat, rice, Triticale, wheat, wheat, rye, barley, ray grass (illustrated with COS-5598).

Figure S6. *De novo* COS sequencing in wheat.

Figure S7. HomoeoSNP vs. SNP calling through sequence coverage criterion.

Figure S8. Characterization of wheat homoeologs (A, B and D homoeoalleles).

Figure S9. Identification of PAVs and CNVs.

Figure S10. COS-SNP and InDels typology in wheat.

Figure S11. Genome-wide distribution of the COS-SNP diversity in wheat.

Figure S12. False COS-SNP origins.

Figure S13. Validation of COS computed gene order.

Figure S14. Detailed characteristics of the WCGM 2013.

Figure S15. Wheat Synteny Viewer characterization flow chart.

Data S1. Experimental procedure.

REFERENCES

Abrouk, M., Zhang, R., Murat, F., Li, A., Pont, C., Mao, L. and Salse, J. (2012) Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole genome duplication. *Plant Cell*, **24**, 1776–1792.

Akhunov, E.D., Akhunova, A.R., Linkiewicz, A.M. *et al.* (2003) Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proc Natl Acad Sci USA*, **100**, 10836–10841.

Akhunov, E.D., Akhunova, A.R., Anderson, O.D. *et al.* (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics*, **11**, 702.

Allen, A.M., Barker, G.L., Berry, S.T. *et al.* (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **9**, 1086–1099.

Allen, A.M., Barker, G.L., Wilkinson, P. *et al.* (2013) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **11**, 279–295.

Bekaert, M., Edger, P.P., Pires, J.C. and Conant, G.C. (2011) Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell*, **23**, 1719–1728.

Berkman, P.J., Visendi, P., Lee, H.C. *et al.* (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol. J.* **11**, 564–571.

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

Cavanagh, C.R., Chao, S., Wang, S. *et al.* (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci USA*, **110**, 8057–8062.

Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M. and Dubcovsky, J. (2009) Analysis of gene-derived SNP marker polymorphism in wheat (*Triticum aestivum* L.). *Mol. Breeding*, **23**, 23–33.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775–780.

Devos, K.M., Dubcovsky, J., Dvorák, J., Chinoy, C.N. and Gale, M.D. (1995) Structural evolution of wheat chromosomes 4A, 5A and 7B and its impact on recombination. *Theor. Appl. Genet.* **91**, 282–288.

Dibari, B., Murat, F., Chosson, A. *et al.* (2012) Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. *BMC Genomics*, **13**, 221.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S. and Wendel, J.F. (2008) Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461.

Duran, C., Edwards, D. and Batley, J. (2009) Genetic maps and the use of synteny. *Methods Mol. Biol.* **513**, 41–55.

Dvorak, J. and Akhunov, E.D. (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops–Triticum* alliance. *Genetics*, **171**, 323–332.

Dvorak, J. and Zhang, H.B. (1990) Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. *Proc Natl Acad Sci USA*, **87**, 9640–9644.

Dvorak, J., Zhang, H.B., Kota, R.S. and Lassner, M. (1989) Organization and evolution of the 5S ribosomal RNA gene family in wheat and related species. *Genome*, **32**, 1003–1016.

Dvorak, J., Akhunov, E.D., Akhunov, A.R., Deal, K.R. and Luo, M.C. (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol. Biol. Evol.* **23**, 1386–1396.

Edger, P.P. and Pires, J.C. (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**, 699–717.

Feldman, M. (1966a) Identification of unpaired chromosomes in F1 hybrids involving *Triticum aestivum* and *T. timopheevii*. *Can. J. Genet. Cytol.* **8**, 144–151.

Feldman, M. (1966b) The mechanism regulating pairing in *Triticum timopheevii*. *Wheat Inf. Serv.* **21**, 1–2.

Feldman, M. and Levy, A.A. (2012b) Genome evolution due to allopolyploidization in wheat. *Genetics*, **192**, 763–774.

Feldman, M., Lupton, F.G.H. and Miller, T.E. (1995) Wheats. In *Evolution of Crop Plants*, 2nd edn (Smartt, J. and Simmonds, N.W., eds). Harlow: Longman Scientific & Technical, pp. 184–192.

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

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.

Furuta, Y., Nishikawa, K. and Yamaguchi, S. (1986) Nuclear DNA content in diploid wheat and its relatives in relation to the phylogeny of tetraploid wheat. *Jpn. J. Genet.* **61**, 97–105.

Gill, B.S. and Chen, P.D. (1987) Role of cytoplasm specific introgression in the evolution of the polyploid wheats. *Proc Natl Acad Sci USA*, **84**, 6800–6804.

Hutchinson, J., Miller, T.E., Jahier, J. and Shepherd, K.W. (1982) Comparison of the chromosomes of *Triticum timopheevii* with related wheats using the techniques of C-banding and in situ hybridization. *Theor. Appl. Genet.* **64**, 31–40.

International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

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.

Jiang, J. and Gill, B.S. (1994) Different species-specific chromosome translocations in *Triticum timopheevii* and *T. turgidum* support the diphyletic origin of polyploid wheats. *Chromosome Res.* **2**, 59–64.

Kilian, B., Ozkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W. and Salamini, F. (2007) Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes. *Mol. Biol. Evol.* **24**, 217–227.

Kota, R.S. and Dvorak, J. (1988) Genomic instability in wheat induced by chromosome 6b(s) of *Triticum speltoides*. *Genetics*, **120**, 1085–1094.

Lai, K., Duran, C., Berkman, P.J. *et al.* (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol. J.* **10**, 743–749.

Lerceteau-Köhler, E., Moing, A., Guérin, G., Renaud, C., Petit, A., Rothan, C. and Denoyes, B. (2012) Genetic dissection of fruit quality traits in the octoploid cultivated strawberry highlights the role of homoeo-QTL in their control. *Theor. Appl. Genet.* **124**, 1059–1077.

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

Liu, Y.G. and Tsunewaki, K. (1991) Restriction fragment length polymorphism (RFLP) analysis in wheat. II. Linkage maps of the RFLP sites in common wheat. *Jpn. J. Genet.* **66**, 617–633.

Lucas, S.J., Akpınar, B.A., Kantar, M. *et al.* (2013) Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS ONE*, **8**, e59542.

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 USA*, **110**, 7940–7945.

Maestra, B. and Naranjo, T. (1999) Structural chromosome differentiation between *Triticum timopheevii* and *T. turgidum* and *T. aestivum*. *Theor. Appl. Genet.* **98**, 744–750.

Murat, F., Xu, J.H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., Messing, J. and Salse, J. (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557.

Naranjo, T. (1990) Chromosome structure of durum wheat. *Theor. Appl. Genet.* **79**, 397–400.

Naranjo, T., Roca, A., Goicoechea, P.G. and Giráldez, R. (1987) Arm homoeology of wheat and rye chromosomes. *Genome*, **29**, 873–882.

Ozkan, H. and Feldman, M. (2001b) Genotypic variation in tetraploid wheat affecting homoeologous pairing in hybrids with *Aegilops peregrina*. *Genome*, **44**, 1000–1006.

Ozkan, H., Levy, A.A. and Feldman, M. (2001a) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops–Triticum*) group. *Plant Cell*, **13**, 1735–1747.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pont, C., Murat, F., Confolent, C., Balzergue, S. and Salse, J. (2011) RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.). *Genome Biol.* **12**, R119.

Quraishi U.M., Abrouk, M., Bolot, S. *et al.* (2009) Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. *Funct. Integr. Genomics*, **9**, 473–484.

Quraishi, U.M., Murat, F., Abrouk, M. *et al.* (2011a) Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (*Triticum aestivum* L.). *Funct. Integr. Genomics*, **11**, 71–83.

Quraishi, U.M., Abrouk, M., Murat, F. *et al.* (2011b) Cross-genome map--based dissection of a nitrogen use efficiency ortho-metaQTL in bread wheat unravels concerted cereal genome evolution. *Plant J.* **65**, 745–756.

Ren, J., Sun, D., Chen, L. *et al.* (2013) Genetic diversity revealed by single nucleotide polymorphism markers in a worldwide germplasm collection of durum wheat. *Int. J. Mol. Sci.* **14**, 7061–7088.

Roder, M.S., Korzun, V., Gill, B.S. and Ganal, M.W. (1998a) The physical mapping of microsatellite markers in wheat. *Genome*, **41**, 278–283.

Roder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. (1998b) A microsatellite map of wheat. *Genetics*, **149**, 2007–2023.

Rong, J., Feltus, F.A., Waghmare, V.N. *et al.* (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics*, **176**, 2577–2588.

Saintenac, C., Jiang, D. and Akhunov, E.D. (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* **12**, R88.

Salina, E.A., Lim, K.Y., Badaeva, E.D., Shcherban, A.B., Adonina, I.G., Amosova, A.V., Samatadze, T.E., Vatolina, T.Y., Zoshchuk, S.A. and Leitch, A.R. (2006) Phylogenetic reconstruction of *Aegilops* section *Sitopsis* and the evolution of tandem repeats in the diploids and derived wheat polyploids. *Genome*, **49**, 1023–1035.

Salse, J. (2012) *In silico* archeogenomics unveils modern plant genome organization, regulation and evolution. *Curr. Opin. Plant Biol.* **15**, 122–130.

Salse, J., Chagué, V., Bolot, S. *et al.* (2008) New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics*, **9**, 555.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schnable, J.C., Freeling, M. and Lyons, E. (2012a) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.* **4**, 265–277.

Schnable, J.C., Wang, X., Pires, J.C. and Freeling, M. (2012b) Escape from preferential retention following repeated whole genome duplications in plants. *Front Plant Sci.* **3**, 94.

Terachi, T., Ogihara, Y. and Tsunewaki, K. (1990) The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. 7. Restriction endonuclease analysis of mitochondrial DNA from polyploid wheats and their ancestral species. *Theor. Appl. Genet.* **80**, 366–373.

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.

Thuillet, A.C., Bataillon, T., Poirier, S., Santoni, S. and David, J.L. (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics*, **169**, 1589–1599.

Trebbi, D., Maccaferri, M., de Heer, P., Sørensen, A., Giuliani, S., Salvi, S., Sanguineti, M.C., Massi, A., van der Vossen, E.A. and Tuberosa, R. (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.). *Theor. Appl. Genet.* **123**, 555–569.

Trick, M., Adamski, N.M., Mugford, S.G., Jiang, C.C., Febrer, M. and Uauy, C. (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol.* **12**, 14.

Wang, X., Shi, X., Hao, B., Ge, S. and Luo, J. (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* **165**, 937–946.

Wang, J., Luo, M.C., Chen, Z., You, F.M., Wei, Y., Zheng, Y. and Dvorak, J. (2013) *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytol.* **198**, 925–937.

Wicker, T., Mayer, K.F., Gundlach, H. *et al.* (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell*, **23**, 1706–1718.

Winfield, M.O., Wilkinson, P.A., Allen, A.M. *et al.* (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.

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 homologs. *PLoS Biol.* **8**, e1000409.

You, F.M., Huo, N., Deal, K.R., Gu, Y.Q., Luo, M.C., McGuire, P.E., Dvorak, J. and Anderson, O.D. (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, **12**, 59.

Zhang, H., Bian, Y., Gou, X. *et al.* (2013) Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. *Proc Natl Acad Sci USA*, **110**, 3447–3452.

Zohary, D. and Feldman, M. (1962) Hybridization between amphidiploids and the evolution of polyploids in the wheat (*Aegilops–Triticum*) group. *Evolution*, **16**, 44–61.

3. Discussion

3.1. L'utilisation des espèces apparentées pour étudier la diploïdisation structurale du génome du blé tendre, un exemple de recherche translationnelle.

Dans le cadre des travaux précédents, publiés en 2013, les séquences parcellaires des génomes des blés diploïdes, tétraploïdes et hexaploïdes n'étaient pas disponibles (*cf.* Figure 6, page 9). Nous avons ainsi mis en œuvre une approche de génomique translationnelle, consistant à utiliser les connaissances disponibles chez les espèces apparentées (les céréales) pour étudier l'organisation du génome du blé tendre. Dans ce contexte, l'étude de la synténie avec des espèces séquencée évolutivement proches, facilite la localisation théorique des gènes chez une espèce non séquencée comme le blé. Toutefois, deux zones synténiques n'étant pas deux copies conformes et donc, pas identiques dans leur contenu en gènes, l'utilisation de plusieurs génomes modèles, optimise la reconstruction du contenu et de l'ordre des gènes chez le blé. On parle dans ce contexte de synténome.

Les travaux de recherche fondamentale en paléogénomique (étude de l'évolution des espèces par la reconstruction des génomes ancestraux disparus) réalisés au sein de l'équipe ont permis de caractériser précisément la synténie chez les céréales (Bolot *et al.* 2009; Salse *et al.* 2008, Murat *et al.* 2014). La paléo-tétraploïdie (WGD) ancestrale des céréales a été identifiée, tout d'abord à partir de la comparaison du génome du riz (Paterson *et al.* 2004; IRGSP, 2005) et des cartes génétiques chez le blé (Salse *et al.* 2008), puis à l'aide des génomes du sorgho (Paterson *et al.* 2009) et du maïs (Schnable *et al.* 2009). Ainsi, après 90 millions d'années d'évolution (date de la 1ère WGD des céréales identifiée), la conservation des gènes ohnologues entre les céréales est de l'ordre de 40% (Bolot *et al.* 2009). Ces travaux ont permis la caractérisation des ancêtres AGK porteur de 7 et 12 chromosomes, à partir duquel, tous les génomes de céréales modernes ont divergé. Les génomes de céréales modernes peuvent être décomposés en 12 (post-paléo-tétraploïdie) ou 7 (pré-paléo-tétraploïdie) blocs ancestraux, permettant de mettre à jour la représentation de la synténie des céréales sous forme des cercles concentriques proposés initialement par M. Gale (sur la base de la comparaison exclusive de cartes génétiques). Dans cette nouvelle représentation (*cf.* Figure 26), les 7 couleurs utilisées représentent les chromosomes ancestraux et les gènes conservés sont reliés entre les espèces (cercles) par des traits gris. En recherche translationnelle, ces gènes conservés peuvent être utilisés pour développer des marqueurs moléculaires (on parle de marqueurs COS pour 'Conserved Orthologous Set') permettant chez une espèce d'intérêt comme le blé, d'affiner la cartographie d'une région de façon robuste, fiable et rapide.

Figure 26. Illustration de la synténie des céréales sous forme de cercles concentriques.
Chaque cercle représente un génome de céréale et les chromosomes (nomenclature des chromosomes mentionnés sur les cercles) conservés sont matérialisés par les couleurs différentes représentant leur origine par rapport à l'ancêtre commun modélisé (A1 – A12) se situant au centre. Les gènes conservés entre cercles concentriques sont reliés par des traits gris. Les flèches noires représentent les réarrangements chromosomiques lors de la transition entre les ancêtres AGK à 7 et 12 chromosomes. *Source Murat et al. 2014.*

Ces travaux ont permis de caractériser 17 317 gènes conservés (considérés comme ancestraux) chez les céréales (Murat *et al*. 2014). J'ai utilisé cette ressource dans le cadre de l'article précédent pour étudier l'organisation du génome du blé par une double approche. La première approche consiste à étudier la rétention de ces gènes ancestraux sur les sous-génomes du blé tendre afin de calculer leur rétention différentielle. Toutefois, cette stratégie se heurte à la possible non-exhaustivité des gènes de blé disponibles. Une autre stratégie consistera donc, en complément, à séquencer les copies homéologues des marqueurs COS afin d'étudier la rétention différentielle des copies homéologues.

Lors de la première stratégie, la carte synténique de 17 317 gènes a été comparée à la séquence génomique parcellaire du blé à disposition en 2012, constituée de 18 483 gènes assignés aux 21 chromosomes. Cette comparaison permet de déceler un biais de conservation des gènes ancestraux entre les sous-génomes. En effet, 35% des gènes ancestraux (conservés chez les céréales) apparaissent présents chez le blé tendre suggérant un taux élevé de gènes 'blé-spécifiques'. Alors que le biais attendu entre les régions D et S, héritées de la paléo-tétraploïdisation, est effectivement observé (au même titre que pour les autres céréales (*cf*. Figure 17, page 19)), un biais de rétention de gènes ancestraux est également retrouvé en réponse à la néo-hexaploïdisation, c'est à dire entre les sous-génomes A, B et D. Toutefois, ce résultat peut être impacté par le génome de référence du blé tendre utilisé ici, qui ne présente pas le catalogue complet de gènes.

Afin de valider cette première observation, la deuxième stratégie consistant à séquencer (après capture) les copies homéologues de marqueurs COS (parmi les 17 317 gènes conservés chez les céréales apparentées) a été mise en œuvre. Elle prend ainsi en considération, un lot de 5 234 gènes COS pour séquencer les 3 homéologues simultanément et sans biais. Ces marqueurs COS sont sélectionnés sur les EST blé disponibles dans les bases de données. Dans mon étude, en l'absence des séquences complètes des 21 chromosomes, le positionnement de ces gènes passera par la cartographie sur une carte génétique consensus, *via* les SNP identifiés. Pour la recherche de polymorphisme, il est alors possible soit de dessiner des primers PCR de part et d'autre des introns (source de polymorphisme), soit de dessiner des sondes permettant la capture et l'enrichissement de ces régions (Figure 27). La recherche de SNP se fait par la suite *via* le séquençage NGS sur divers cultivars ; par PCR-seq ou Capture-seq (Figure 27).

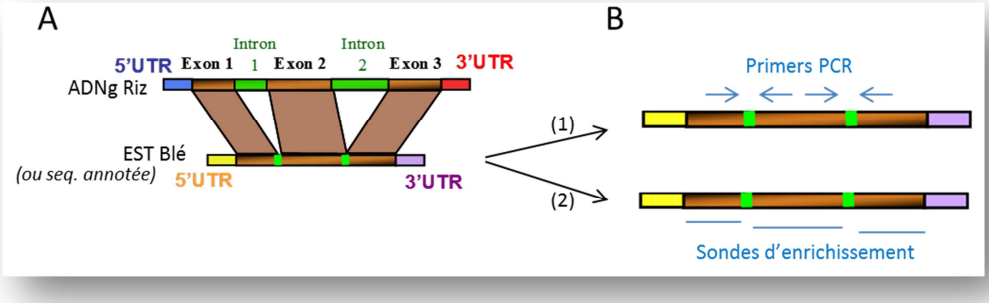


Figure 27. Design de marqueurs COS par sélection d'amorces PCR ou de sondes d'enrichissement.
(A) Illustration schématique d'un marqueur COS identifié par comparaison d'une séquence génomique de riz (ADNg) et d'une EST de blé. Les bornes introns (vert) / exons (marron) sont positionnées sur la séquence de blé par alignement des séquences, dans un second temps les sondes et primers peuvent être sélectionnés sur cette séquence. (B) Design des amorces PCR à partir de la séquence blé, amplifiant les introns (1) ou des sondes (2) hybridant sur les exons.

Dans ce contexte, à partir de la carte synténique de 17 317 gènes, 5 234 marqueurs COS ont été développés. Le séquençage (chez la variété *Chinese Spring*) des 5 234 marqueurs a délivré 23 463 gènes blé potentiels qui ont été capturés et séquencés chez 8 variétés, délivrant un catalogue de 9 969 SNPs. Cette ressource nous a permis de caractériser les différences structurales en termes de nombre de marqueurs COS (parmi les 5 234), nombre de SNPs (parmi les 9 969), ratio CM/marqueur (à partir des 807 SNPs cartographiés) et ceci pour les compartiments génomiques de la dominance chez le blé tendre, hérités des 3 événements de polyploïdisation. Ainsi, une évolution contrastée entre les compartiments expliquerait l'asymétrie structurale observée entre les sous-génomes en termes de contenu en gènes, structure de la chromatine et du polymorphisme des gènes (marqueurs). Ainsi, la néo-dominance se caractérise par la compartimentation des génomes homéologues où le sous-génome B apparaît sensible (plastique) par rapport aux sous-génomes A et D, dits dominants, en réponse à la tétraploïdisation puis à l'hexaploïdie.

Toutefois, l'asymétrie observée entre les sous-génomes du blé (avec du plus plastique au plus stable B>A>D) peut être de diverses origines (*cf.* Figure 28). Dans l'article précédent, j'ai ainsi proposé un scénario évolutif du génome du blé avec une dominance des sous-génomes post-polyploïdie à l'origine de cette asymétrie structurale. Cette asymétrie observée entre les sous-génomes A, B et D peut être de 3 origines.

(1) L'asymétrie observée entre les-sous génomes A, B et D du blé tendre hexaploïde, peut être héritée d'une asymétrie préexistante entre les trois progéniteurs *T. urartu*, *A. speltoides et A. tauschii* (*cf.* Figure 28, effet progéniteur). Notamment, il est possible que l'ancêtre géniteur B soit inconnu, car disparu. Bien que proche d'*A. speltoides*, ce progéniteur B pourrait être porteur d'une importante variabilité structurale (comparativement aux ancêtres des sous-génomes A et D) et ainsi porteur d'une plasticité pré-polyploïdisation observée sur le sous-génome B moderne (Talbert *et al*. 1995 ; Blake *et al*. 1999).

(2) L'asymétrie observée entre les sous-génomes A, B et D du blé tendre hexaploïde, peut également être héritée d'événements d'hybridations entre les différents progéniteurs ou entre différents génotypes de chacun de ces progéniteurs, pré-polyploïdisation. On parle notamment d'origine polyphylétique du sous-génome B (Blatter *et al*. 2004) et d'une origine hybride homoploïde du sous-génome D. Dans ce dernier cas, l'événement d'hybridation entre les progéniteurs A et B aurait donné le sous-génome D il y a 5 MYA (*cf.* Figure 28, effet hybridation ; Marcussen *et al*. 2014; Sandve *et al.* 2015), et pourrait impacter l'asymétrie observée entre les sous-génomes du blé hexaploïde moderne.

(3) Enfin, l'asymétrie observée entre les sous génomes A, B et D du blé tendre hexaploïde, peut être héritée d'une plasticité post-polyploïdisation générée par l'évolution différentielle (dominance) des sous-génomes. Cette origine consisterait à la diploïdisation structurale non aléatoire des sous-génomes post-polyploïdie qui pourrait expliquer l'asymétrie observée entre les sous-génomes du blé moderne, comme je l'ai proposé dans l'article précédent (*cf.* Figure 28, effet dominance).

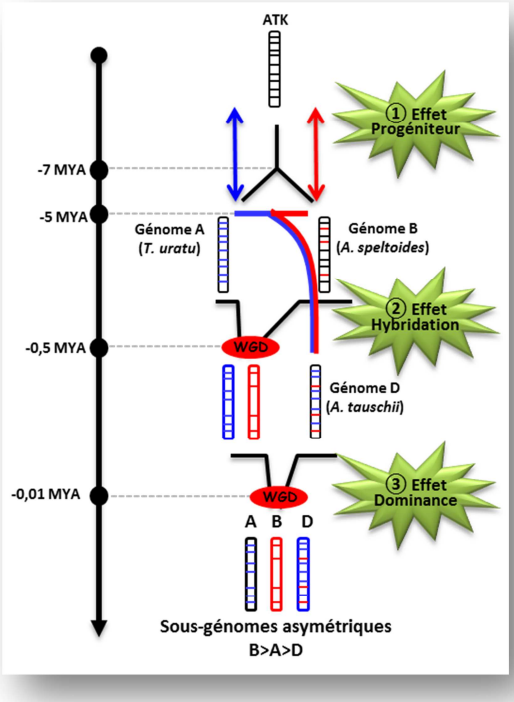


Figure 28. Scénario évolutif à l'origine de l'asymétrie structurale des sous-génomes du blé hexaploïde.
Illustration de l'asymétrie observée entre les sous-génomes A, B et D du blé tendre hexaploïde à partir de 3 scénarios : (1) divergence structurale entre les progéniteurs, (2) hybridation entre les progéniteurs et (3) dominance des sous-génomes. Dans ce dernier cas, suite à l'évènement de tétraploïdisation le sous-génome A issu de *T. urartu* est dit dominant (matérialisé en bleu), celui issue d'A. *speltoides* (B) est dit sensible (matérialisé en rouge). Le dernier évènement de polyploïdisation s'additionne et engendre l'asymétrie des sous-génomes du blé hexaploïde avec une plasticité structurale B>A>D. La datation des évènements de polyploïdie (d'après Jordan *et al.* 2015) est référencée sur l'échelle à gauche de la figure.

3.2. Asymétrie des sous-génomes par dominance post-polyploïdie

Notre étude publiée en 2013 dans 'Plant Journal' a permis de suggérer pour la première fois, qu'il existe des différences structurales (et génétiques) au sein du génome moderne du blé tendre, héritées de la duplication ancestrale (il y a ≈90 MYA) et de l'hexaploïdie récente (<0.5 MYA). Cette évolution contrastée entre les compartiments hérités des événements de polyploïdie, expliquerait l'asymétrie structurale observée entre les sous-génomes en termes de contenu en gènes, structure de la chromatine et du polymorphisme des gènes. Nous avons ainsi, proposé un modèle original de dominance des sous-génomes post-polyploïdie chez le blé moderne, additionnant les effets des deux polyploïdisations récentes, avec l'effet de la dominance ancestrale. Concernant la paléo-dominance, elle serait engagée chez le blé suite à l'évènement WGD il y a 90 MYA, au même titre que pour toutes les céréales issues d'AGK, où la dominance est déjà établie (Murat *et al.* 2014). Concernant la néo-dominance, elle serait engagée suite à l'évènement de tétraploïdisation (il y a 500 000 ans), entre le sous-génome A issu de T. *urartu* dit dominant, matérialisé en bleu (*cf.* Figure 29), et le sous-génome B issu d'A. *speltoides* dit sensible (matérialisé en orange). Dans notre modèle, nous proposons que lors du dernier évènement de polyploïdisation (il y a 10 000 ans), le sous-génome D issu de A. *taushii* dit dominant, matérialisé en bleu (cf. Figure 29), fusionne avec le tétraploïde dit sensible (matérialisé en orange). Nous avons vu que la plasticité issue de la néo-dominance, observable au sein des génomes homéologues, interagit avec la paléo-dominance générant un contraste en 6 compartiments (3 sous-génomes homéologues x 2 paléo-sous-génomes ; S-A, S-B, S-D, D-A, D-B et D-D ; *cf.* Figure 29). A titre d'exemple, je montre dans mon étude de 2013, que le chromosome 1B est supra-S (double sensibilité) car il est issu du chromosome ancestral 1 pré-ATK qui est Sensible (S) et localisé sur le sous-génome B (Sensible lors des deux événements de polyploïdie du blé). Sur ce chromosome, la transposition et la perte de gènes ancestraux est forte, ainsi que le polymorphisme SNP et l'accumulation d'hétérochromatine révélés par C-banding. A l'opposé, le chromosome 3D est décrit comme supra-D (double dominance) car il est issu du chromosome ancestral 3 pré-ATK qui est Dominant (D) et localisé sur le sous-génome D (Dominant lors du deuxième événement de polyploïdie du blé), et à ce titre il se caractérise de la façon suivante : stable par la rétention des gènes ancestraux et peu de transpositions, peu de polymorphismes et d'hétérochromatine. Au-delà des chromosomes ou fragments chromosomiques supra-D (3D, 2D-S/L, 4D, 5D-L, 6D-S/L) et supra-S (1B, 2B-C, 5B-S, 6B-C, 7B), certains groupes chromosomiques, tels que T2, T5 et T6 (*cf.* Figure 29) sont issus de la fusion de régions D et S, et excluent donc l'appellation supra-D ou supra-S. A titre d'exemple, les chromosomes (A, B et D) du groupe 2, sont issus de la fusion de chromosomes ancêtre AGK A7(S) et A4(D). Le chromosome 2B ne peut être dit supra-S car il possède en lui une région dominante ancestral (A4) et sensible (sous-génome B) en réponse à la néo-hexaploïdie.

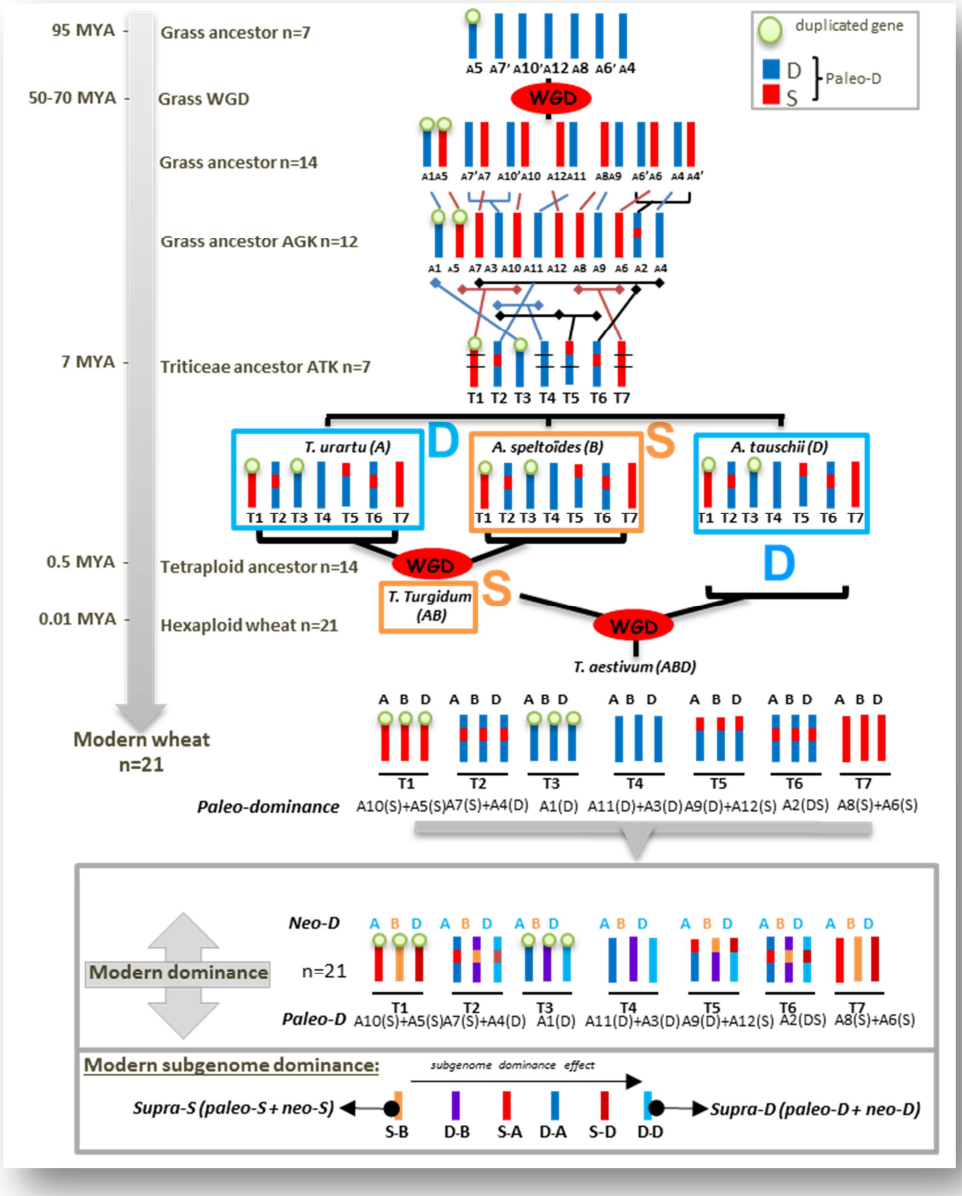


Figure 29. Dominance des sous-génomes du blé tendre hexaploïde.

L'histoire évolutive du blé est schématisée à partir de l'ancêtre des céréales AGK à 7 chromosomes. Cet ancêtre a subi une duplication totale de son génome doublant le contenu chromosomique puis deux réarrangements pour aboutir à un ancêtre à 12 chromosomes. Cette WGD a engendré un effet de dominance des deux sous-génomes, matérialisé ici en bleu, pour la fraction dominante (A1-2-3-4-9-11) et en rouge, pour la fraction sensible (A5-6-7-8-10-12). Cette compartimentation peut être identifiée sur les 21 chromosomes du blé en tenant compte des 5 fusions chromosomiques (transition AGK → ATK) par translation de la synténie. A titre d'exemple, un gène ancestral matérialisé par un cercle vert, présent en une copie dans l'ancêtre AGK est potentiellement présent en 6 copies chez le blé moderne. Les relations d'orthologie entre les 7 chromosomes ancestraux (A1 à A12) et les 21 chromosomes du blé tendre (T1 à T7 avec les trois sous-génomes A, B et D) sont mentionnées au bas des chromosomes de blé. Ainsi la paléo-dominance et paléo-sensibilité des sous-génomes du blé tendre apparaissent comme suit T1(S)=A10(S)+A5(S), T2(D/S)=A7(S)+A4(D), T3(D)=A1(D), T4(D)=A11(D)+A3(D), T5(D/S)=A9(D)+A12(S), T6(D/S)=A2(D/S), T7(S)=A8(S)+A6(S). La néo-dominance des sous-génomes est matérialisée par des rectangles bleus sur les sous-génomes dominants A et D, et orange pour B sensible. La double dominance intégrant la paléo et néo-dominance chez le blé moderne hexaploïde est modélisée dans le rectangle gris en bas de la figure. Il est possible de visualiser les 6 compartiments S-A, S-B, S-D, D-A, D-B et D-D représentés par 6 couleurs différentes. Les blocs dits supra-dominants (supra-D c'est-à-dire D-D) et supra-sensibles (supra-S c'est-à-dire S-B) sont matérialisés en bleu clair et en orange respectivement. Paleo-D: Paléo-dominance; Neo-D: Néo-dominance; AGK : ancestral grass karyotype ; ATK : ancestral *Triticeae* karyotype ; WGD : whole genome duplication. La datation des évènements est indiquée en millions d'années (MYA).

Ce modèle de plasticité structurale contrastée entre sous-génomes chez le blé tendre, est renforcé par l'étude de Zhang *et al.* 2013 qui a étudié les réarrangements chromosomiques de plus de 1 000 blés synthétiques (c'est-à-dire néo-polyploïde) par hybridation *in situ*. Les auteurs concluent que sur les 3 génomes, le B donne plus fréquemment lieu à l'aneuploïdie (anomalie du lot chromosomique), suivi de A. Le génome D est largement le plus stable ne présentant aucun réarrangement dans ce contexte expérimental.

3.3. L'apport des nouvelles séquences génomiques

Description des données et de la méthodologie **-** Le séquençage complet du génome du blé hexaploïde et de ses ancêtres diploïdes réalisé ces dernières années (*cf.* Figure 6 page 9), ouvre de nouvelles perspectives quant à la compréhension de l'organisation des gènes chez le blé tendre. A partir des séquences du blé hexaploïde disponibles, il est possible de réactualiser les données produites dans le cadre de l'article précédent et ainsi, d'affiner nos connaissances sur la dominance structurale des sous-génomes. En effet, en 2014, a été publiée une nouvelle référence du génome du blé (IWGSC 2014), issue du séquençage WGS des bras de chromosomes. Cette nouvelle référence comporte 99 386 gènes assignés aux bras des 21 chromosomes (Borrill *et al.* 2015). L'alignement blast de ces gènes et l'application de seuils d'alignements (identité cumulée de 90 % et longueur d'alignement cumulée de 30%, comme décrits dans Salse *et al.* 2009), ont permis avec le programme MCL (http://micans.org/mcl/), d'identifier les triplets de gènes homéologues A, B et D sur les 21 chromosomes.

Dans un second temps, à partir de ces 99 386 gènes, Florent Murat (bioinformaticien de l'équipe) a réactualisé le synténome du blé en utilisant le génome ancestral des céréales et ainsi, ordonner 72 900 gènes sur les 21 chromosomes. Ce synténome est l'intégration des données de génomiques (scaffolds blé), de génétiques (marqueurs moléculaires) et de paléogénomiques (gènes AGK) par alignement des séquences de ces ressources comme décrit dans l'article Plant Journal 2013 et illustré dans la Figure 30.

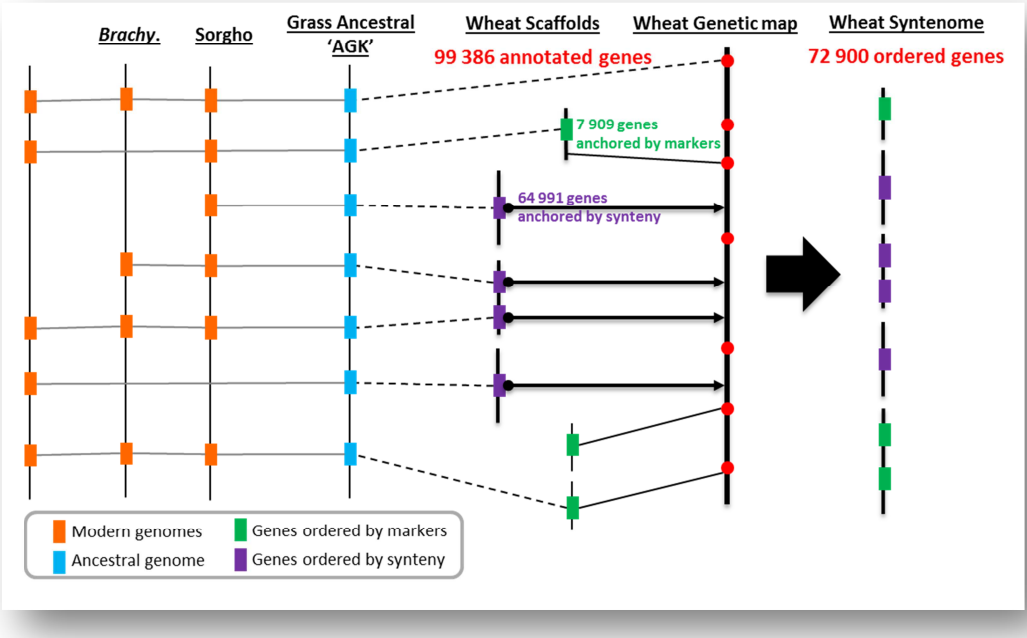


Figure 30 : Représentation de la construction du « Synténome » chez le blé
A gauche, sont représentés en orange, les gènes des espèces modèles (riz, *Brachypodium* et sorgho) et leurs conservations permettant de reconstruire l'ancêtre des céréales AGK, en bleu, composé de 12 chromosomes. Les 99 386 séquences génomiques du blé sont ancrées sur les chromosomes par alignement avec les marqueurs moléculaires produits par Wang *et al.* 2014 (points rouges), soit 7 909 gènes ancrés par les marqueurs (en vert). Les scaffolds ne pouvant être positionnés sur la carte génétique sont ancrés par synténie *via* l'ancêtre des céréales AGK, soit 64 991 gènes conservés en violet. Le synténome ainsi produit est constitué de 72 900 gènes positionnés sur les 21 chromosomes du génome du blé tendre. AGK : Grass Ancestral Karyotype. *Source : F Murat.*

L'analyse 'Gene Ontology' (GO) de ces données, permet d'étudier l'enrichissement selon les catégories de gènes avec une nomenclature normalisée. La base de données GO des 99 386 gènes du blé est disponible sur agriGO (http://bioinfo.cau.edu.cn/agriGO/). L'analyse statistique et la visualisation sont réalisées avec les outils agriGO et revigo (http://revigo.irb.hr/) en filtrant les gènes avec un FDR<0.05 et p-value p<.001.

Résultats **-** L'alignement blast de 99 386 gènes ont permis d'identifier 8 671 triplets strictement homéologues. Ces 8 671 triplets correspondent à 26 % des gènes (*cf.* Figure 31A), mais de nombreux gènes possèdent au moins une copie supplémentaire. En effet, 6 673 clusters de gènes possèdent au moins 4 copies ou plus (35 971 gènes, représentant 36% du répertoire total ; *cf.* Figure 31A, cercles oranges). Au total, 62% des gènes sont présents au minimum en 3 copies homéologues. Au-delà des triplets, nous avons identifié 5 157 paires d'homéologues (représentant 10% du répertoire total) avec une perte potentielle d'une copie homéologue. Les gènes retrouvés en singleton (15 761) montrent une diploïdisation totale de l'ordre de 16 % des gènes (2 copies homéologues perdues). En prenant en compte les clusters où une à deux copies ont été perdues (ou pouvant être localisées sur un chromosome non-homéologue, *cf.* Figure 31A, cercles oranges) une diploïdisation est engagée pour 38 % des gènes au total chez le blé tendre.

Les résultats de l'analyse GO montrent que pour les 15 761 singletons (A, B ou D), 31 catégories sont enrichies pour des processus biologiques, notamment de type 'DNA replication' et 'RNA-dependent DNA replication' (*cf.* Figure 31B). La diploïdisation s'est donc opérée pour les processus biologiques de réplication de l'ADN, et ce, dans de très fortes proportions par rapport à l'ensemble du génome. En effet, 34% des gènes de la catégorie 'RNA-dependent DNA replication' sont présents en singletons (et 17% concernant la catégorie 'DNA réplication'). Une seule des 3 copies initiales suffit à assurer ces fonctions. Lorsque l'on compare les singletons des sous-génomes ABD entre eux, un enrichissement significatif (p<.001) est retrouvé sur chacun des sous-génomes pour ces mêmes catégories.
Concernant les 8 671 triplets (*cf.* Figure 31C), les résultats montrent que 160 GO sont enrichies, et les plus significatives (p<.001) relèvent des processus biologiques de 'regulation of cellular process' et 'biological regulation'. Pour ces catégories, les trois copies homéologues sont préférentiellement retenues au cours de l'évolution. La spécificité de chacune des copies homéologues permet potentiellement de garantir la fréquence, la vitesse ou l'étendue de ces processus biologiques, soit une certaine plasticité.

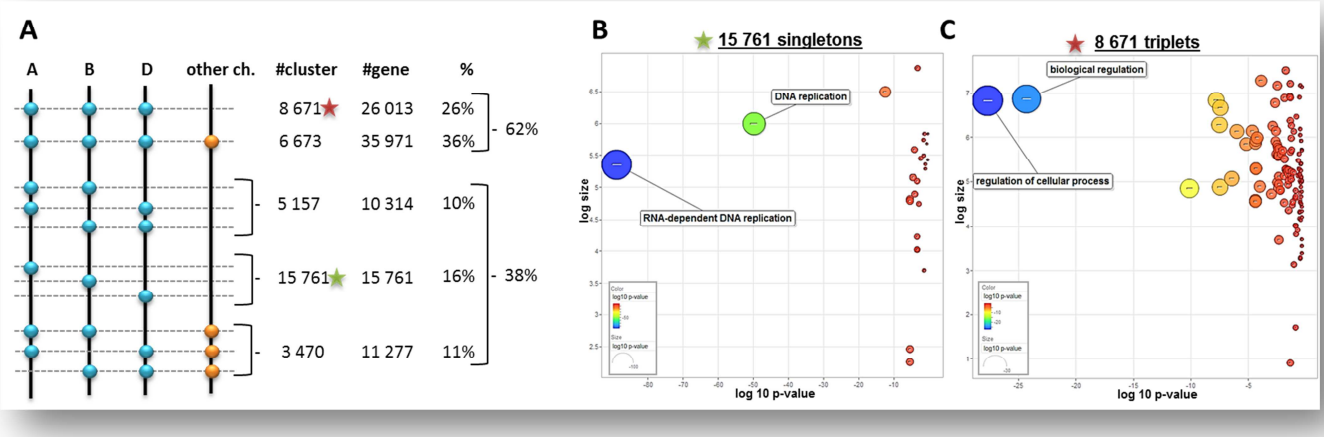
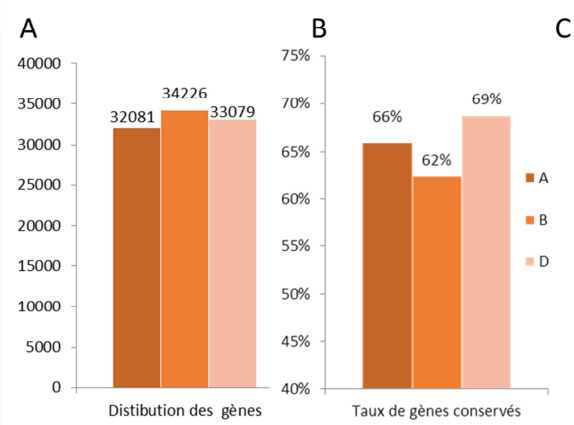


Figure 31 : Distribution structurale des 99 386 gènes du blé tendre hexaploïde et description ontologique.
(A) Répartition des 99 386 gènes du blé tendre hexaploïde disponibles en 2014. La répartition structurale des gènes homéologues identifiés est schématisée à gauche sur des chromosomes A, B et D par des cercles bleus. Certains gènes sont dupliqués sur d'autres chromosomes ('other ch.') et sont représentés en orange. Les effectifs des clusters et des gènes pour chaque catégorie sont mentionnés dans la partie droite avec le % (par rapport au répertoire génique total de 99 386 gènes). Au regard de la perte des copies homéologues, une diploïdisation est engagée pour 37,6 % des gènes. (B) et (C) : Enrichissement GO (Gene Ontology) selon l'ontologie des processus biologiques pour les 15 731 singletons (B) et les 8 671 triplets (C). L'analyse statistique et la visualisation sont réalisées avec les outils agriGO, et revigo en filtrant les gènes avec un FDR<0,05. L'axe des abscisses représente le log 10 de la p-value et l'axe des ordonnées indique la fréquence du terme GO dans la base de données blé. La couleur et la taille des cercles sont proportionnelles à la p-value, selon l'échelle en bas à gauche.

J'ai pu actualiser, à partir de ces nouvelles séquences, notre connaissance de l'asymétrie structurale des sous-génomes du blé tendre. A partir des 99 386 gènes disponibles (Bolser *et al.* 2014), l'analyse de la distribution des gènes orthologues (gènes ancestraux conservés) confirment les résultats établis dans l'article Pont *et al.* 2013 à l'échelle du génome entier, avec les sous-génomes du plus plastique (perte de gènes orthologues) au plus stable (rétention des gènes orthologues) : B<A<D. En effet, le contenu génique global est assez homogène, voire légèrement plus élevé sur le sous-génome B (avec 32 081, 34 226 et 33 079 gènes respectivement pour les génomes A, B et D, *cf.* Figure 32 A), si on considère tous les gènes annotés et assignés aux 21 chromosomes. Cependant, si on considère uniquement les gènes ancestraux conservés chez le blé tendre (contenus dans le synténome de 72 900 gènes ordonnés sur les 21 chromosomes), l'analyse montre une perte de gènes (*cf.* Figure 32B) au sein du génome sensible B (avec 62,3 % d'orthologie versus 65,9 % et 68,7 % pour respectivement les sous-génomes A et D). Cette nouvelle ressource confirme le biais de rétention des gènes ancestraux conservés chez les céréales apparentées entre les sous-génomes B<A<D. Au niveau de la GO, il existe un enrichissement significatif (p<.001 ; par rapport au contenu génique global du blé) sur les sous-génomes B et D (*cf.* Figure 32C) pour leur contenu en gènes associé aux catégories GO suivantes : 'photosynthesis' (spécifiquement pour le sous-génome D), 'DNA packaging' ou encore 'cell wall macromolecule catabolic process' (spécifiquement pour le sous-génome B). Si l'on considère les gènes de blé orthologues aux céréales apparentées et donc conservés après 90 MYA d'évolution, ceux-ci sont enrichis en GO correspondant à des gènes essentiels à la vie de la cellule; 'response to oxidative stress', 'response to chemical' et 'oxidation-reduction process'. On retrouve également des catégories qui correspondent au maintien d'un d'équilibre interne de la cellule (cellular homeostasis ; retrouvé ancestralement sur les sous-génomes A et B) ou à la biosynthèse de macromolécules constitutives de paroi cellulaire ('cell wall organisation or biogenesis' sur B et D).



GO number	genomes	GOA	P-value	FDR
GO:0015979	D	photosynthesis	6,3E-08	6,9E-05
GO:0071669	B	plant-type cell wall organization or biogenesis	1,7E-06	9,8E-04
GO:0065004	B	protein-DNA complex assembly	8,1E-06	4,0E-03
GO:0016998	B	cell wall macromolecule catabolic process	1,3E-05	5,9E-03
GO:0006323	B	DNA packaging	2,8E-05	1,1E-02
GO:0042592	A & B	homeostatic process	5,3E-05	1,8E-02
GO:0034728	A & B	nucleosome organization	1,8E-06	9,8E-04
GO:0045454	A & B	cell redox homeostasis	5,8E-05	1,8E-02
GO:0006333	A & B	chromatin assembly or disassembly	1,5E-06	9,8E-04
GO:0019725	A & B	cellular homeostasis	3,8E-05	1,4E-02
GO:0071554	B & D	cell wall organization or biogenesis	1,0E 08	1,4E 05
GO:0006412	A & D	translation	3,3E-05	3,0E-02
GO:0042221	A, B & D	response to chemical stimulus	1,0E-13	2,8E-10
GO:0006979	A, B & D	response to oxidative stress	1,5E-20	7,9E-17
GO:0055114	A, B & D	oxidation reduction	1,3E-09	2,3E-06

Figure 32. Distribution des gènes et GO associée au sein des sous-génomes modernes A, B et D.
(A) Distribution totale en gènes (99 386 gènes) des sous-génomes modernes A, B et D. (B) Proportion (%) de gènes conservés chez les céréales (orthologues contenus dans le synténome constitué de 72 900 gènes) rapportés aux 99 386 gènes du répertoire génique disponible. (C) Enrichissement GO (Gene Ontology) des génomes homéologues par rapport au synténome *via* agriGO. Les valeurs de p-value et FDR sont mentionnées dans les colonnes de droite.

Dans un second temps, à partir des 72 900 gènes du synténome, j'ai pu attribuer à chacun des gènes ancestraux ordonnés sur les 21 chromosomes, un des 6 compartiments de dominance (S-A ; D-A ; S-B; D-B; S-A; S-D). La distribution des gènes orthologues, confirme nos premières observations de perte massive de gènes sur les régions paléo-sensibles (38 708 gènes sur la fraction paléo-D et 26 283 sur la fraction paléo-S ; *cf*. Figure 33A). La néo-dominance relevant de la néo-polyploïdie (500 000 et 10 000 ans) est moins marquée car plus récente, ce temps évolutif n'étant pas assez long pour permettre la délétion complète de gènes dans ce contexte. Pour avoir plus de poids dans les analyses statistiques des asymétries entre A, B et D, il faudrait plus d'observations et pour cela, analyser d'autres génotypes de blé tendre avec une même résolution de séquençage. J'ai également pu calculer le taux de conservation des gènes orthologues (64 991 gènes du synténome, *cf*. Figure 30, gènes violets). J'ai ainsi observé une rétention de gène graduelle (Figure 33B) du compartiment le plus stable au plus plastique : D-D>S-D>D-A >S-A >D-B>S-B. Avec ces nouvelles données, plus exhaustives, je retrouve aux extrémités de cette distribution (reflétant la plasticité) les compartiments dits ultra-sensibles du génome B (S-B * en rouge, avec 85,7 % de gènes retenus sur le synténome) et supra-dominants des chromosomes D (D-D * en bleu, avec 93,4 % de gènes retenus), matérialisés respectivement en orange clair et bleu clair dans la Figure 29.

Pour conclure, une rétention de gènes non-aléatoire est observée entre les différents compartiments génomiques hérités des événements de polyploïdisation chez le blé tendre. Un effet de diploïdisation significatif est observé au sein du génome sensible B et de rétention au sein du génome D (Figure 33B, à droite).

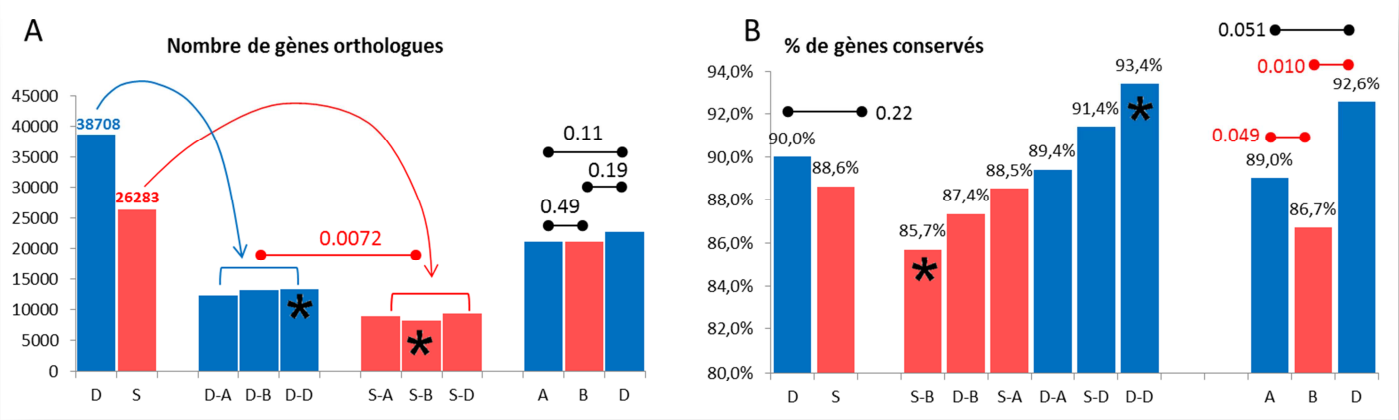


Figure 33. Rétention différentielle en gènes des sous-génomes du blé tendre.
Distribution de la compartimentation différentielle des gènes. Les groupes dit sensibles sont représentés en rouge, et dominants en bleu. Les groupes supra-D et supra-S sont matérialisés par un astérisque. Les p-values sont mentionnées au-dessus des histogrammes (T-test unilatéral par paires), en rouge les valeurs significatives. Les histogrammes correspondent aux compartiments de paléo-dominance (à gauche, S et D) ; néo-dominance (à droite ; A, B et D) et l'intégration des 6 compartiments (3 sous-génomes homéologues x 2 paléo-sous-génomes ; au centre ; S-A, S-B, S-D, D-A, D-B et D-D) en réponse aux paléo-duplications et néo-duplications. (A) Distribution des gènes ancestraux *via* l'effectif total en gènes conservés dans chaque compartiment (99 386 gènes) et contenus dans le synténome (72 900 gènes). (B) Proportion (en %) de gènes ancestraux dans chaque compartiment, rapporté aux 64 991 gènes orthologues du synténome.

L'analyse GO montre un enrichissement différentiel significatif (p<.001) entre les différents compartiments supra-D et supra-S (*cf.* Figure 34). Le compartiment supra-S, à forte diploïdisation, est enrichi en gènes des catégories GO 'response to chemical' et 'response to oxidative stress' notamment. Ces processus se traduisent par un changement d'état ou d'activité d'une cellule (en termes de mouvement, sécrétion, production d'enzyme, d'expression des gènes, etc., en réponse aux stress oxydatifs ou à la suite d'un stimulus chimique. Quant au compartiment supra-D, à forte stabilité structurale, il est enrichi en GO, sans doute essentielles et constitutives (car ancestrales) à la vie de la cellule telles que 'cell wall organisation' pour la biosynthèse de macromolécules constitutives de paroi cellulaire ou encore pour la photosynthèse. On retrouve ici les mêmes catégories que lors de la GO des gènes ancestraux (*cf.* Page 41), ce qui est attendu, car le sous-génome D-D montre le plus fort taux de rétention en gènes ancestraux. Les résultats GO des fonctions moléculaires montrent que la fraction sensible du génome du blé est impliquée dans la réponse aux stimuli, et ce, *via* l'interaction spécifique entre molécules (catégorie 'binding'). Au contraire les gènes portés par la fraction dominante du génome du blé semblent être essentiels aux fonctions cellulaires de base avec des fonctions moléculaires très constitutives ('heme binding', 'oxidoreductase activity', 'fructose-bisphosphate aldolase activity'…).

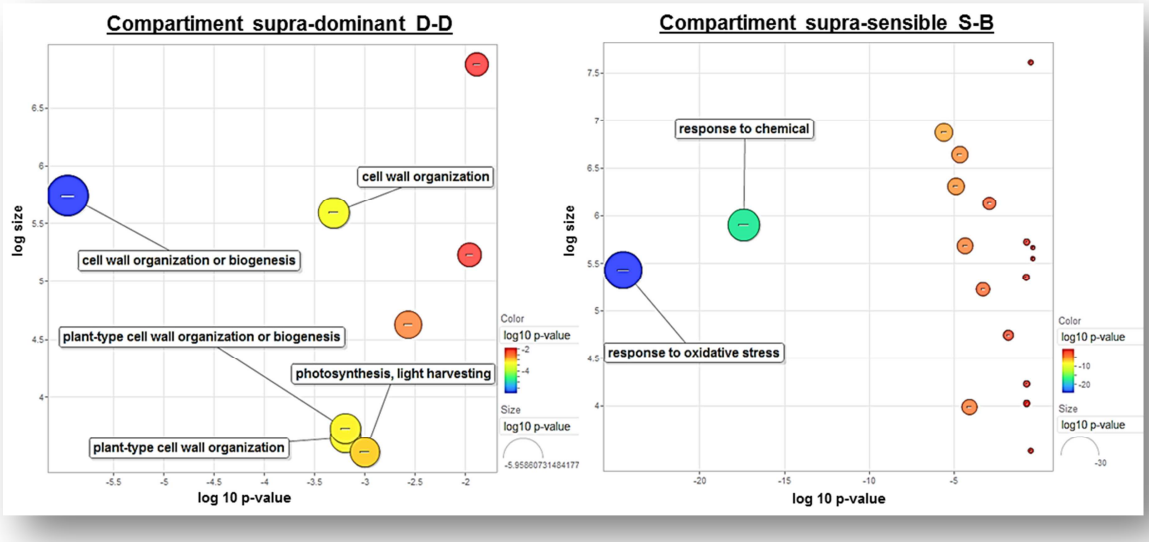


Figure 34 : description ontologique des processus biologiques associés aux compartiments supra-D et supra-S.
Enrichissement en gènes des compartiments supra-D (à gauche) et supra-S (à droite) selon l'ontologie des processus biologiques. L'analyse statistique et la visualisation sont réalisées avec les outils agriGO, et revigo en filtrant les gènes avec un FDR<0,05. L'axe des abscisses représente le log 10 de la p-value et l'axe des ordonnées indique la fréquence du terme GO dans la base de données blé. La couleur et la taille des cercles sont proportionnelles à la p-value, selon l'échelle en bas à droite.

3.4. Asymétrie structurale *via* les homéoSNPs

Les travaux précédents, que ce soit à partir du séquençage de marqueurs COS (article Pont *et al.* 2013), ou à partir de la dernière version du génome du blé (chapitre précédent), ont été réalisés en comparant les compartiments chromosomiques du blé hexaploïde. L'asymétrie observée, que nous proposons être le fruit du processus de dominance des sous-génomes, peut tout ou en partie, être héritée de la divergence structurale des génomes des trois progéniteurs (*Triticum urartu, Aegilops speltoides* et *Aegilops tauschii*). Quelle part de l'asymétrie structurale des sous-génomes du blé tendre moderne est imputable au progéniteur et/ou à la dominance des sous-génomes ? Enfin, cette asymétrie structurale nous renseigne-t-elle sur l'origine controversée du blé hexaploïde par rapport aux trois progéniteurs : *Triticum urartu, Aegilops speltoides* et *Aegilops tauschii ?*

La séquence du génome du blé hexaploïde, disponible depuis 2014, composée de 99 386 gènes (IWGSC 2014, Borrill *et al.* 2015, Bolser *et al.* 2014), associée à celles des géniteurs diploïdes (Jia *et al.* 2013, Luo *et al.* 2013, Ling *et al.* 2013) permet d'étudier la plasticité structurale du blé hexaploïde héritée, soit des progéniteurs ou soit des événements de polyploïdisation. Deux approches ont été menées dans ce contexte. La première consiste à recenser les gènes présents chez les progéniteurs qui seraient perdus au cours du temps chez le blé hexaploïde, et donc uniquement imputable aux événements de polyploïdie. Toutefois, ces assemblages demeurent parcellaires et de tels gènes non identifiés chez l'hexaploïde pourraient être dus à un problème de séquençage (ou d'assemblage) de génomes. J'ai ainsi proposé dans le cadre du travail initié par Moaine Elbaidouri (Post-doctorant de l'équipe), de travailler l'évolution du génome du blé tendre non par l'analyse de la perte de gènes, mais par une analyse détaillée des mutations entre sous-génomes (les homéoSNPs) et leurs héritabilités à partir des progéniteurs. A partir des 8 671 triplets homéologues chez le blé hexaploïde, nous avons pu identifier

précisément les homéoSNPs partagés entre homéologues (*cf*. Figure 35), et donc ceux spécifiques des sous-génomes A (en vert), B (en rouge) et D (en bleu).

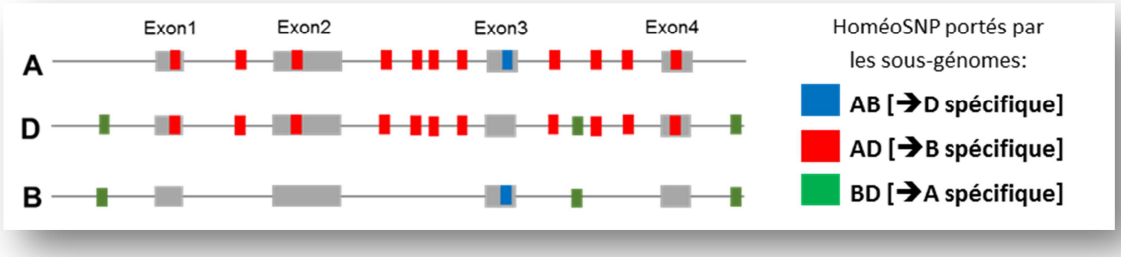


Figure 35. Identification des homéoSNPs
Cette figure représente de façon schématique les 3 copies homéologues A, B et D d'un gène composé de 4 exons (rectangles gris). Les homéoSNPs sont matérialisés en rouge lorsqu'ils sont communs entre les copies A et D (les homéoSNPs sont alors 'B-spécifiques'), en vert, ceux communs entre B et D (les homéoSNPs sont alors 'A-spécifiques'), et bleu, les homéoSNPs partagés entre A et B (les homéoSNPs sont alors 'D-spécifiques'). *Source : Moaine El Baidouri*

On comprend de façon assez intuitive que l'on peut, par cette méthode, étudier l'apparentement des gènes homéologues. Ce type d'analyse a été réalisé par Marcusen *et al.* 2014, avec une approche classique de phylogénie, et a permis justement d'identifier l'hybridation homoploïde du sous-génome D, provenant de l'hybridation des progéniteurs A et B. Ce résultat a été très controversé (Li et *al.* 2015ab, Sandve *et al.* 2015) et n'est pas encore totalement admis dans la communauté scientifique. L'apport de nouvelles séquences génomiques permet également d'étudier plus finement cette question. Toutefois l'origine controversée du sous-génome D, n'entre pas directement dans l'étude de la dominance post-polyploïdie. Elle sera développée à la fin de ce chapitre, participant à la compréhension de l'évolution du génome du blé et illustrant l'asymétrie structurale pré-polyploïdie.

Plasticité différentielle (dominance) post-polyploïdie - Est-ce que l'analyse de l'évolution du profil de mutations nous permet de déceler une plasticité différentielle des sous-génomes post-polyploïdie ? C'est-à-dire, la dominance des sous-génomes ? A partir des 8 671 triplets de gènes homéologues chez le blé hexaploïde, nous avons pu identifier 3 121 triplets pour lesquels des séquences orthologues chez les diploïdes et le tétraploïde sont identifiées, permettant ainsi la comparaison simultanée des 6 copies de gènes (3 copies du blé hexaploïde et 3 copies parentales, *cf*. Figure 36A-C). Pour chacun des sous-génomes nous avons recensé les homéoSNPs hérités du progéniteur et ceux absents du progéniteur en question. Pour le génome A, les homéoSNPs spécifiques de ce sous-génome (mutations non présentes sur les deux autres homéologues) sont à 71% retrouvées chez le progéniteur *T. urartu (cf*. Figure 36C). De la même manière, pour le génome D, les homéoSNPs spécifiques de ce sous-génome sont à 75% retrouvées chez le progéniteur *A. tauschii.* Mais, lorsque l'on regarde l'héritabilité des homéoSNPs identifiés sur le sous-génome B à partir du progéniteur *A. speltoides*, ce pourcentage chute à 42% *(cf*. Figure 36B). En effet, 54% des homéoSNPs 'B-spécifiques' du blé moderne ne sont pas transmis du progéniteur *A. speltoides*. Cette observation peut être expliquée par les arguments classiquement invoqués dans la littérature : une origine plus ancienne d'*A. speltoides*, par rapport aux deux autres progéniteurs (Talbert *et al*. 1995 ; Blake *et al*. 1999) ou une origine polyphylétique (Blatter *et al*. 2004).

Dans cette dernière possibilité, le génome B serait le produit de l'hybridation de plusieurs ancêtres disparus aujourd'hui dont *A. speltoides* ne serait qu'un parent éloigné, *A. speltoides* n'étant donc pas l'ancêtre proprement dit du sous-génome B moderne.

En introduisant dans cette analyse le blé tétraploïde (*T. durum* AB), il devient possible d'identifier, parmi les homéoSNPs du sous-génome B du blé hexaploïde non transmis à partir du progéniteur B, ceux présents chez le tétraploïde (donc perdus seulement après la tétraploïdie) et ceux absents chez le tétraploïde (donc perdus avant la tétraploïdie). Sur cette base, pour les homéoSNPs du sous-génome B du blé hexaploïde non transmis du progéniteur B, 84% sont présents chez le tétraploïde mais 16% apparaissent spécifiques de l'hexaploïde *(cf*. Figure 36D). Par comparaison pour le sous-génome A, de la même manière, 61% des homéoSNPs sont présents chez le tétraploïde et 39% apparaissent spécifiques de l'hexaploïde. Nous avons ici caractérisé des mutations (homéoSNPs) qui sont apparues spécifiquement après les événements de polyploïdie (entre les états 2X et 4X mais aussi entre 4X et 6X).

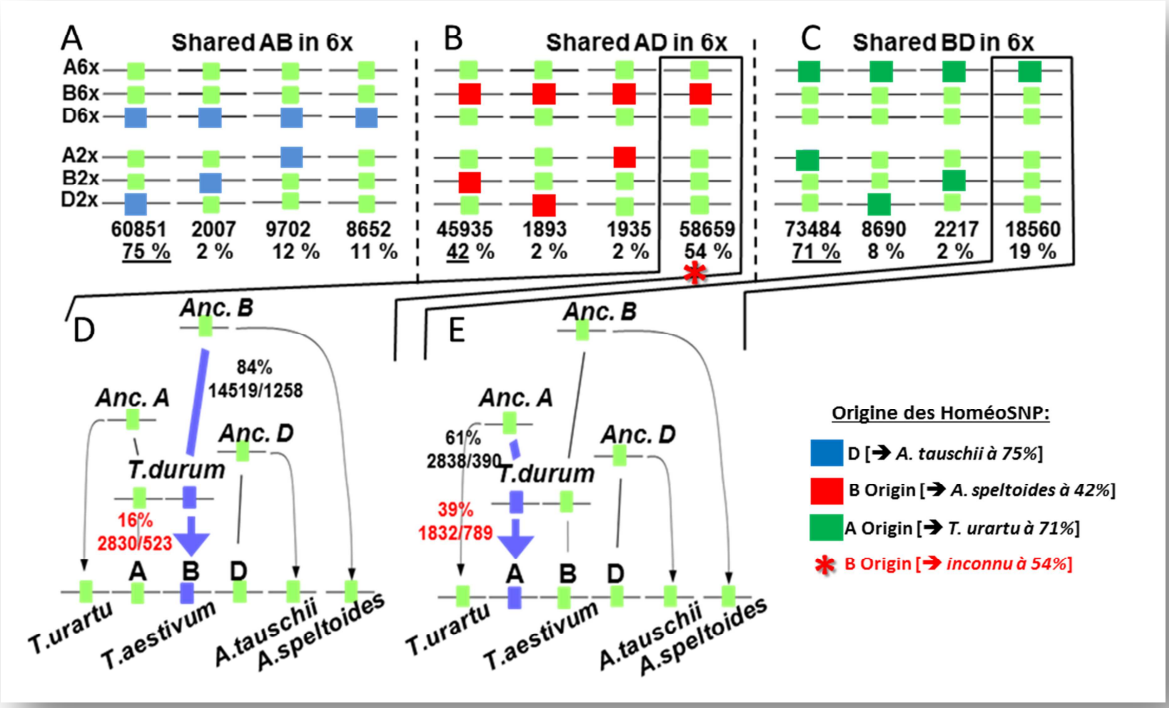


Figure 36. HoméoSNPs partagés entre homéologues du blé moderne et ses progéniteurs

En haut de la figure sont matérialisées les différentes copies orthologues chez le blé hexaploïde (6x), et les progéniteurs diploïdes (2x) avec les homéoSNPs représentés par les carrés colorés. Les homéoSNPs sont classés en 3 cas de gauche à droite : (A) les homéoSNPs partagés entre A et B et provenant donc du sous-génome D, (B) les homéoSNPs partagés entre A et D et provenant donc du sous-génome B et (C) les homéoSNPs partagés entre D et B et provenant donc du sous-génome A. (D) Description des homéoSNPs 'B-spécifiques' du blé moderne non conservés avec A. *speltoides* (E) et les homéoSNPs 'A-spécifiques' non conservés avec T. *uratu via* l'illustration du modèle évolutif supportant. Les lignes bleues illustrent les différents cas de transmission des homéoSNPs. Le % est indiqué et en dessous, le nombre d'homéoSNPs apparus avec le nombre de gènes impliqués à partir des parents diploïdes et du tétraploïde T. *durum* et enfin l'hexaploïde (T. *aestivum*). Chez l'hexaploïde, 84% des homéoSNPs 'B-spécifiques' ne sont pas transmis par A. *speltoides* confirmant l'origine polyphylétique (A. *speltoides* n'est pas l'ancêtre proprement dit). *Adapté de El Baidouri et al. 2016.*

A-t-on une accélération de l'accumulation de mutations entre les sous-génomes A et B lors de la transition diploïde/tétraploïde et tétraploïde/hexaploïde ? Concernant le sous-génome A, nous avons identifié 7,3 homéoSNPs /gènes (2 838 homéoSNPs pour 390 gènes, *cf*. Figure 37B) apparus entre le diploïde (*T. urartu*) et le tétraploïde ainsi que 2,3 homéoSNPs /gènes (1 832 homéoSNPs sur 789 gènes, *cf*. Figure 37B) apparus entre le tétraploïde et l'hexaploïde. De la même manière (*cf*. Figure 37A), nous avons recensé pour le sous-génome B, 11,5 homéoSNPs /gènes (14 519 homéoSNPs sur 1258 gènes) apparus entre le diploïde (*A. speltoides*) et le tétraploïde, ainsi que 5.4 homéoSNPs /gènes (*i.e.* 2 830 homéoSNPs sur 523 gènes) apparus entre le tétraploïde et l'hexaploïde. La comparaison de ces taux montre que le passage entre les états diploïdes et tétraploïdes a engendré l'apparition de 1,5x (11,5/7,3) fois plus de mutations sur le sous-génome B par rapport au sous-génome A (*cf*. Figure 37). Le passage entre les états tétraploïdes et hexaploïdes a engendré l'apparition de 2,3x (5,4/2,3) fois plus de mutations sur le sous-génome B par rapport au sous-génome A. Le génome B accumule plus de mutations que les autres sous-génomes en réponse aux deux évènements de polyploidïsation. Toutefois les gènes accumulant des mutations entre tétraploïdes et hexaploïdes ne montrent pas d'enrichissement significatif (p<.05) au niveau de la GO. Cette accélération de plasticité (ici au travers de l'accumulation de mutations) valide le modèle évolutif du blé tendre par dominance des sous-génomes en réponse à la polyploïdisation, proposée dans l'article de ce chapitre, Pont *et al.* 2013.

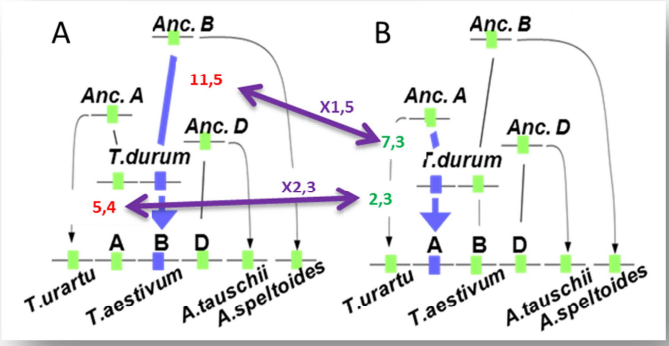


Figure 37. Nombre moyen d'homéoSNPs apparus entre progéniteurs diploïdes et tétraploïdes.
(A) Modèle évolutif matérialisant les homéoSNPs 'B-spécifiques' du blé moderne non transmis du progéniteur A. *speltoides*. (B) Modèle évolutif matérialisant les homéoSNPs 'A-spécifiques' du blé moderne non transmis du progéniteur *T. urartu* (B). Les lignes bleues illustrent les différents cas de transmission des homéoSNPs avec le nombre moyen d'homéoSNPs apparus par gène à partir des parents diploïdes vers le tétraploïde T. *durum* et du tétraploïde vers l'hexaploïde (T. *aestivum*). L'accumulation de mutations sur le sous-génome B post-polyploïdie est matérialisée par les flèches violettes. Le passage entre les états tétraploïdes et hexaploïdes a engendré l'apparition de 2,3x fois plus de mutations sur le sous-génome B (en rouge) par rapport au sous-génome A (en vert). Sur sous-génome B, 11,5 homéoSNPs /gènes (14 519 homéoSNPs sur 1258 gènes ; *cf.* Figure 36) sont apparus entre le diploïde (A. speltoides) et le tétraploïde *Adapté de El Baidouri et al. 2016.*

Plasticité différentielle pré-polyploïdie, origines du blé réconciliées - A partir des séquences génomiques (gènes et introns inclus, *cf*. Figure 35 plus haut) des 8 671 triplets homéologues, nous avons pu calculer précisément le taux d'homéoSNPs partagés par gène sur chaque homéologues (*i.e.* 'apparentement' ou 'proximité' des copies entre AB ; AD et BD). Basée sur l'apparentement des gènes, Marcussen *et al.* 2014 dans la revue science, ont suggéré une hybridation homoploïde entre les géniteurs diploïdes A et B,

donnant naissance au génome D avec une contribution égale. Cette conclusion provient d'une analyse phylogénétique de triplets de gènes homéologues montrant que les gènes à fort apparentement B-D (A[B/D]) et A-D (B[A/D]) sont deux fois plus abondants que les gènes à fort apparentement A-B (D[A/B]) malgré une spéciation entre D-A et D-B plus récente que celle de A-B (respectivement ≈5,5 MYA et ≈6,5 MYA ; Marcussen *et al.* 2014). Li *et al.* 2015 ont repris une partie des données de Marcussen *et al.* 2014 en incluant d'autres lignées du complexe *Triticum* et d'*Aegilops* en prenant en compte les génomes nucléaires mais aussi chloroplastiques. Ces derniers auteurs montrent qu'*A. Tauschii* (progéniteur D) ne partage pas le génome chloroplastique ni du géniteur A (*T. urartu)*, ni du B *(A. speltoides)* qui serait attendu sur la base d'une origine hybride du progéniteur D proposée par Marcussen *et al.* 2014. Les auteurs suggèrent que l'origine du progéniteur D et donc du sous-génome D du blé hexaploïde est plus complexe qu'une hybridation homoploïde (D=A+B).

Notre analyse, basée sur les homéoSNPs partagés entre les homéologues, montre que l'apparentement des gènes au sein des triplets est de 42 % et 38 % entre les homéologues AD et BD respectivement, mais chute à 20 % entre les sous-génomes A et B (*cf.* Figure 38). Au sein du génome moderne du blé, seulement 20% des homéoSNPs sont D spécifiques.

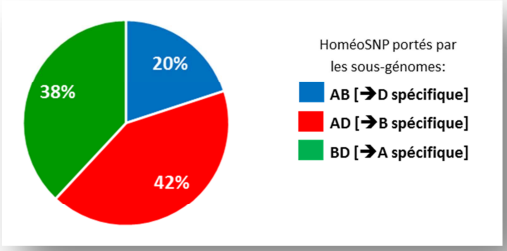


Figure 38. Proximité de séquence entre les copies homéologues.
Résultats de proximité entre les copies A, B et D des 8 671 triplets sur la base des homéoSNPs partagés entre les trois copies de gènes. Au sein du génome moderne du blé, seulement 20% des homéoSNPs sont D spécifiques.

Ce résultat suggère que les gènes des sous-génomes A et D ou B et D sont plus proches entre eux, que ceux des sous-génomes A et B. Ceci peut être observé du fait de l'origine hybride du sous-génome D entre les progéniteurs ancêtres '*Anc* A' (*T. urartu) et 'Anc B' (A. speltoides*), expliquant 80% de la structure du sous-génome D du blé hexaploïde (Marcussen *et al.* 2014). Toutefois 20% de la structure du sous-génome D du blé hexaploïde ne peut provenir de ces deux progéniteurs A et B mais impose d'introduire dans le modèle un troisième progéniteur D (*Anc D').* Ces résultats démontrent l'origine hybride du sous génome D du blé hexaploïde avec une contribution inégale de trois progéniteurs *Anc A, Anc B* et *Anc D.* Notre étude basée sur l'analyse de l'évolution du pattern de mutations (HoméoSNPs) réconcilie les deux études (Li *et al.* 2015, Marcussen *et al.* 2014), en montrant clairement la participation des 3 progéniteurs dans l'architecture d'*A. Tauschii* (et du sous-génome D du blé hexaploïde moderne) mais avec une contribution inégale (*cf.* Figure 39).

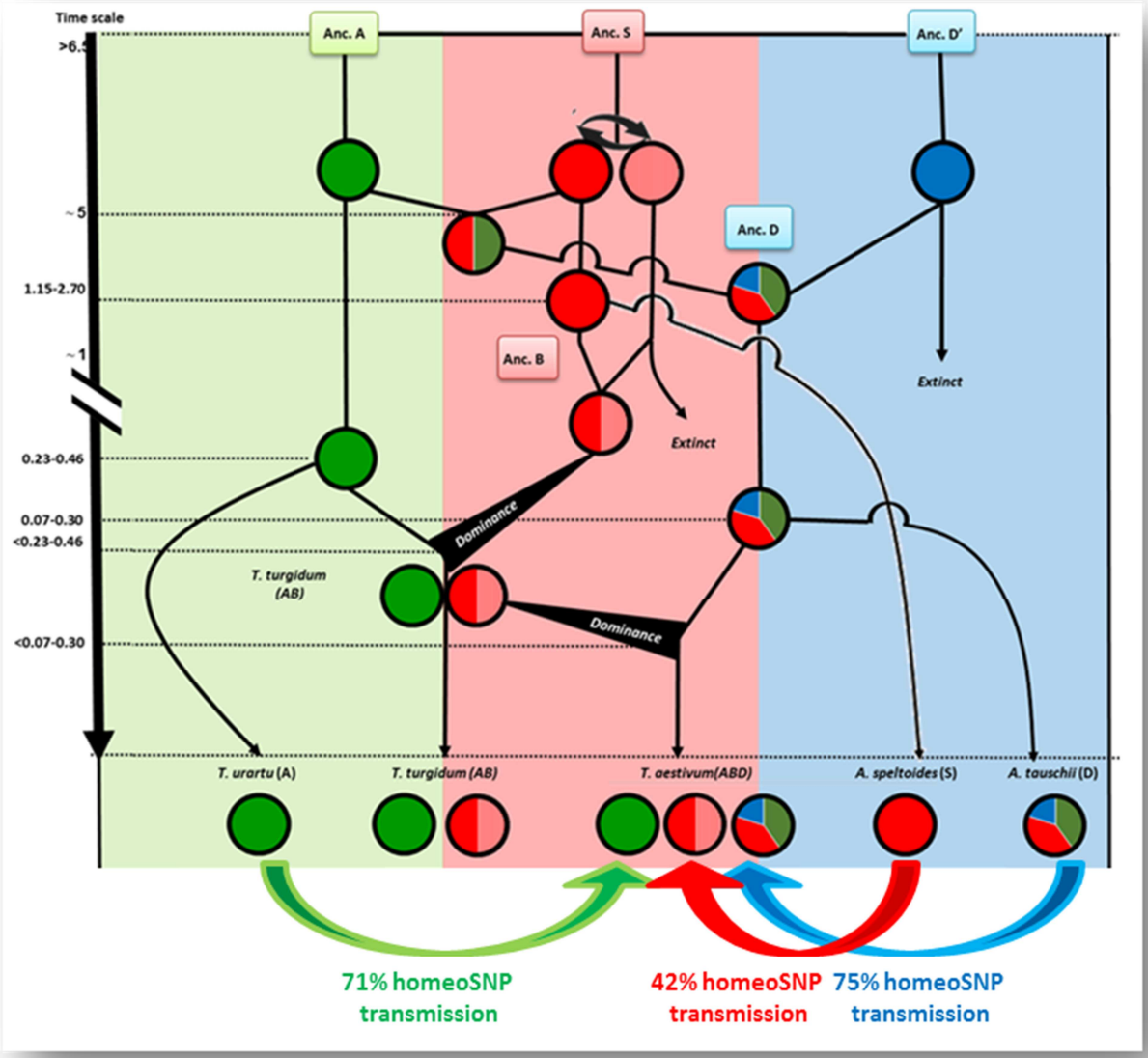


Figure 39. Illustration de l'origine évolutive du blé tendre réconciliée.
Illustration de l'histoire évolutive du blé tendre à partir d'un ancêtre AncA (Ancestor genome A, cercle vert), AncB (Ancestor genome B, cercle rouge) et AncD (Ancestor genome D, cercle bleu). Le temps est exprimé en millions d'années sur la flèche noire à gauche. Les sous-génomes sont illustrés sous forme de cercles et lors des hybridations, sous forme de camemberts colorés matérialisant la contribution de chaque progéniteur. Les espèces modernes sont représentées en bas de la figure. Les pourcentages de mutations transmises par les progéniteurs aux différents sous-génomes du blé hexaploïde moderne sont illustrés au bas de la figure. Le génome moderne A provient de T. *urartu*. Le génome moderne B provient d'un ancêtre proche A. *speltoides* qui cumulerait une origine polyphylétique et un effet de dominance des sous-génomes post-polyploïdisation. Le génome moderne D provient de A. *tauschii* après une hybridation homoploïde et une origine indépendante Anc. D'. *Source : El Baidouri et al. 2016.*

En étudiant l'origine des homéoSNPs du blé tendre, il est ainsi possible d'éclaircir l'origine des gènes du blé tendre sur les 21 chromosomes (*cf.* Figure 40). L'apparentement des gènes du blé hexaploïde peut être matérialisé sur chaque chromosome pour observer un 'patchwork' de 3 couleurs matérialisant

l'hybridation homoploïde du sous-génome D (*cf.* Figure 40, cercle 1, couleurs vertes et rouges sur le génome D) ainsi que l'apport spécifique d'un ancêtre Anc D' (couleur bleu). L'origine polyphylétique du sous-génome B est aussi visible (*cf.* Figure 40 cercle 2, couleurs grises) avec les gènes dont l'origine ne peut être identifiée (ni A. *tauschii*, ni A. *speltoides* et ni T. *urartu)*.

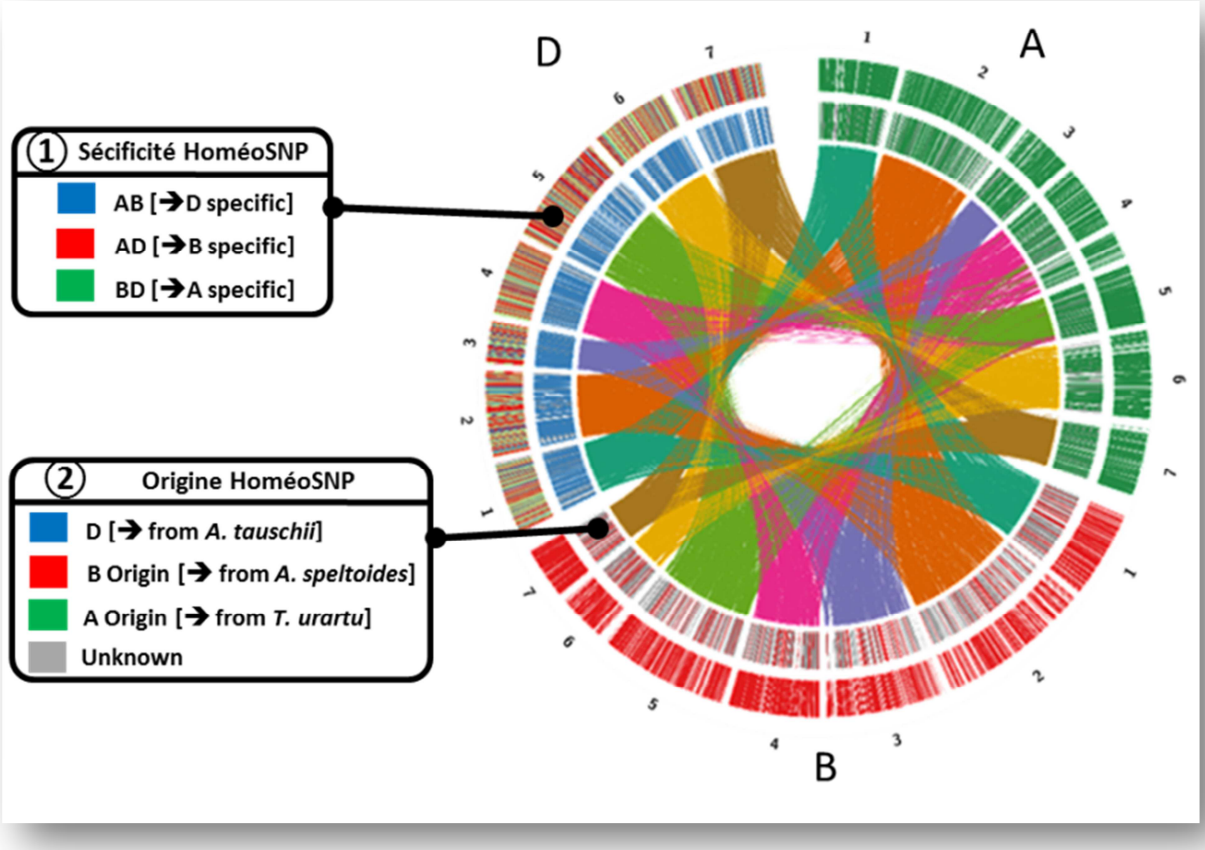


Figure 40. Visualisation de l'asymétrie structurale *via* l'origine des HoméoSNPs du blé tendre.
Illustration de l'apparentement des gènes du blé sur la base des homéoSNPs. Chaque cercle reprend les 21 chromosomes du blé hexaploïde selon l'ordre modélisé des 72 900 gènes du synténome. Au centre sont matérialisées les relations d'homéologies avec les 8 671 triplets A, B et D. Le cercle n°1, le plus externe, représente l'apparentement des gènes selon la conservation des homéoSNPs. L'apparentement BD est en vert, matérialisant les gènes spécifiques du sous-génome A, l'apparentement AD est en rouge, matérialisant les gènes spécifiques du sous-génome B, l'apparentement AB est en bleu matérialisant les gènes spécifiques du sous-génome D. Les génomes modernes A et B sont constitués de gènes spécifiques A et B, alors que le génome D est constitué d'un patchwork de 3 couleurs matérialisant l'hybridation homoploïde AB (couleurs verte et rouge) ainsi que l'apport spécifique d'un ancêtre Anc D' (couleur bleu). Le cercle n°2, illustre l'origine des gènes selon la conservation des homéoSNPs avec les géniteurs diploïdes, A. *tauschii*, A. *speltoides* et T. *urartu* . Les gènes d'origine A. *tauschii* sont matérialisés en bleu, les gènes d'origine A. *speltoides* sont matérialisés en rouge, les gènes d'origine T. *uratu* sont matérialisés en vert et les gènes dont l'origine ne peut être identifiée avec ces trois géniteurs sont matérialisés en gris. L'origine des sous-génomes A et D est sans ambiguïté, mais le sous-génome B possède de nombreux gènes dont l'origine ne peut être identifiée (en gris), reflétant l'origine polyphylétique du sous-génome B et/ou la plasticité de ce sous-génome en réponse aux évènements de polyploïdie (effet de la dominance). *Source : Mélanie Molinier.*

Conclusions du chapitre - L'analyse de la dynamique évolutive des gènes et mutations chez les blés diploïdes, tétraploïdes et hexaploïdes nous permet d'affiner le scénario évolutif du blé moderne où :
(1)Le sous-génome A du blé moderne est hérité à 71% du progéniteur T. *urartu*.
(2)Le sous-génome B du blé moderne est d'origine polyphylétique (impliquant probablement un ancêtre éteint qui est alors inconnu à nos jours), autrement dit, *A. speltoides* n'est pas l'ancêtre proprement dit du sous-génome B moderne avec seulement 42% d'homéoSNPs transmis de *A. speltoides* au sous-génome B du blé moderne. Il existerait donc un ancêtre B ; Anc S, inconnu et proche de *A. speltoides*. Cette plasticité particulière du sous-génome B est également le fruit de la diploïdisation non aléatoire des sous-génomes post-polyploïdie *via* la perte de gènes mais aussi l'accumulation de mutations au cœur des gènes, la dominance des sous-génomes. Ainsi, le phénomène de dominance serait engagé suite à l'évènement d'hexaploïdisation, avec une accumulation accélérée d'homéoSNPs sur le sous-génome B.
(3)Une hybridation homoploïde entre les parents diploïde A et B a donné naissance à un ancêtre Anc D à l'origine de *A. Tauschii*, puis du sous-génome D moderne. Une part non négligeable des gènes du sous-génome D (20 %), provient ni de A et ni de B et donc d'un donneur spécifique (ancêtre Anc D'). Cette conclusion implique donc que le génome D du blé moderne ait connu 2 évènements d'hybridation pour aboutir à *A. Tauschii* (*cf.* Figure 40); un entre les ancêtres Anc. A et Anc. S puis un, avec l'ancêtre Anc. D'.

Ces résultats ont été valorisés à travers un nouvel article, accepté en Juin 2016, dans la 'revue New Phytologist' dont je suis 'co-corresponding' auteur ayant proposé et discuté longuement l'étude de l'évolution du blé tendre par la dynamique d'accumulation des mutations (homéoSNPs) « El Baidouri M, Murat F, Veyssiere M, Molinier M, Pont C*, Salse J*. Reconciliating the evolutionary origin of bread wheat (*Triticum aestivum*) ». L'article El Baidouri *et al.* 2016 est fourni en annexe de ce manuscrit.

4. Perspectives ; l'étude d'ADN ancien de blé

Les résultats que nous avons obtenus sur la plasticité structurale des sous-génomes du blé tendre ont été acquis en comparant les sous-génomes du blé tendre (article Plant journal) avec les représentants modernes des progéniteurs diploïdes et tétraploïdes (réactualisation des résultats). Toutefois la perte de gènes ou l'accumulation de mutations peuvent être liées à la divergence des représentants modernes (diploïdes ou tétraploïdes) à partir des véritables ancêtres disparus. Dans ce contexte la séquence d'ADN de restes archéobotaniques (on parle d'ADN ancien ou aDNA) donnerait un point fixe dans l'évolution passée à partir duquel il pourra être recensé la perte de gènes ou les mutations accumulées chez l'hexaploïde depuis 10 000 ans. A ce titre, avoir accès à des séquences génomiques aDNA, même fragmentées, permettrait de valider le modèle de dominance des sous-génomes. Pour valider l'accélération d'accumulation de mutations du génome B (venant de sa supra-sensibilité à la fois ancienne et récente), mais aussi la rétention des gènes sur les compartiments supra-dominants, l'analyse pourra se porter sur plusieurs échantillonnages de diploïdes, tétraploïdes, hexaploïdes, et ce à des échelles géologiques variables. Des études ont déjà été menées avec des échantillons de blés d'Europe

centrale datant du Néolithique (thèse de Mehmet Somel, 2003). Ces travaux montrent que l'amplification d'aDNA d'échantillons carbonisés n'est pas chose facile, et que les chances de succès sont faibles. Cependant, les technologies permettent maintenant de recueillir et purifier l'aDNA de plantes desséchées, congelées, immergées ou carbonisées (Shapiro 2012). Dans ce cadre, un partenariat est débuté avec l'INRAP (institut national de recherches archéologiques) de Clermont-Ferrand et le muséum d'histoire naturelle à Paris. Ils disposent d'échantillons de restes archéologiques de blé datant de plus 3200 ans avant Jésus-Christ. J'ai débuté au sein de l'équipe un projet d'enrichissement sélectif de l'aDNA par capture pour le séquençage des 16 464 protogènes de l'ancêtre des céréales AGK. Il sera alors possible d'étudier la vitesse de divergence de chaque sous-génome par l'étude de la perte de gènes et d'accumulation de mutations (homéoSNPs) entre individus d'espèces différentes et notamment entre blés tétraploïdes et hexaploïdes, sur 5 000 ans.

5. Conclusion

La dominance des génomes est une réponse à la polyploïdie qui induit un retour compartimenté à la diploïdie structurale du génome post-polyploïdie. Chez le blé moderne hexaploïde, elle a pu être caractérisée et quantifiée ; une diploïdisation est engagée pour 38 % des gènes avec une sensibilité différentielle des sous-génomes B>A>D (avec une conservation contrastée des gènes ancestraux de 62 %<66 %<69 %). La paléo-dominance (en réponse à la duplication ancestrale) se surimpose à la polyploïdie récente (hexaploïdie), pour donner naissance aux compartiments supra-dominant (D-D avec une forte rétention de gènes ancestraux) et supra-sensible (S-B avec une plus faible rétention de gènes ancestraux). Les travaux de cette thèse permettent ainsi de proposer le premier modèle évolutif du blé tendre dans lequel les sous-génomes ont été architecturés à partir de l'hybridation des trois ancêtres mais également à partir de la dominance des sous-génomes post-polyploïdie avec une plasticité accélérée (ou sensibilité) du sous-génome B. La dominance, ou plus généralement l'asymétrie structurale, des sous-génomes du blé tendre, a-t-elle un impact sur la régulation des gènes homéologues ?

CHAPITRE 2

Impact de la polyploïdie sur la régulation du génome du blé tendre

1. Introduction à l'étude de l'impact de la polyploïdie sur la régulation du génome du blé tendre

1.1. Polyploïdie et asymétrie de l'expression des gènes chez les plantes.

Dans ce deuxième volet de mes travaux de thèse, je me suis intéressée à comprendre l'impact d'une duplication globale du génome (WGD) sur la régulation des gènes. Dans la littérature, une modification de l'expression des copies dupliquées a été observée chez de nombreuses espèces (Chang *et al.* 2010, Flagel *et al.* 2009). En effet, chez les polyploïdes, alors qu'on aurait pu s'attendre à un effet redondant des gènes parentaux pré-duplication, les copies redondantes montrent une expression divergente, pouvant aller jusqu'au silencing de l'une des copies. Le terme 'silencing' est retrouvé dans la littérature pour l'extinction de l'expression d'au moins un homéologue. Dans le cadre de travaux chez le coton (AADD) allotétrapolyploïde depuis 1 à 2 millions d'années (MYA), les auteurs (Chaudhary *et al.* 2009 ; Flagel *et al.* 2009) rapportent un biais d'expression favorisant le génome homéologue D dans cinq des espèces polyploïdes étudiées. De même, chez *Arabidopsis suecica (AABB)*, allotétrapolyploïde depuis 12 000 à 300 000 ans, la rétention de gènes associés à un fort niveau d'expression a été démontrée en faveur de son sous-génome homéologue dominant AA (Chang *et al.* 2010). Lors du séquençage récent de *Brassica napus* (AACC), polyploïde depuis 7 500 ans, l'analyse n'a pas mis en évidence de biais d'expression entre les sous-génomes A et C (Chalhoub *et al.* 2014). Cependant au sein même de cette étude, lorsqu'un tissu ou un stade de développement est considéré, il a été montré un déséquilibre d'expression en faveur de certains loci homéologues (mais cependant pas à l'échelle plus large de blocs génomiques voire de sous-génomes). Enfin, les travaux de l'équipe auxquels j'ai participé en 2009, ont permis de quantifier la divergence d'expression des gènes dupliqués chez le riz issus de la duplication ancestrale datant de 90 MYA. Les résultats montrent dans cette étude, que plus de 85 % des gènes dupliqués ont divergé dans leur profil d'expression au cours du développement du grain, mais aussi de la feuille et de la racine (Throude *et al.* 2009).

1.2. Polyploïdie et asymétrie de l'expression des gènes chez le blé.

Chez le blé hexaploïde, certaines études (Bottley *et al.* 2006 ; Mochida *et al.* 2006) suggèrent une forte diploïdisation expressionnelle des copies homéologues (c'est-à-dire une élimination de la redondance d'expression des gènes dupliqués). Dès 2006, une étude de l'asymétrie expressionnelle des gènes homéologues (A, B et D) a été réalisée à partir 236 EST (Bottley *et al.* 2006) indiquant que 27 % des gènes sont partiellement 'silencés' dans la feuille et 26 % dans la racine. D'autres résultats similaires (Mochida *et al.* 2006), confirment que les profils d'expression des copies homéologues chez le blé, varient selon les tissus considérés, où les copies s'expriment alternativement. Dans cette dernière étude, parmi les gènes étudiés au cours du développement du grain, 27 % montrent une préférence d'expression dans un des trois sous-génomes et 81 % montrent une différence d'expression de gènes homéologues dans au moins un des 8 tissus testés. La fréquence de cette variabilité d'expression varie selon le tissu, avec une prédominance de plasticité (spécificité d'expression) observée dans certains tissus très spécialisés comme le pistil (Mochida *et al.* 2006). Plus récemment, *via* le séquençage RNA-seq des transcrits chez le

blé au cours du développement du grain, 28 % des triplets présents au sein du génome du blé tendre ont une expression redondante au cours du développement du grain de blé, selon Pfeifer *et al.* 2014.

Conclusion : la littérature reporte une asymétrie d'expression des copies de gènes dupliqués héritée de duplications totales de génome (anciennes ou récentes) chez les plantes et notamment entre certaines copies homéologues chez le blé tendre. Dans l'article qui suit, paru dans la revue 'Genome Biology' en décembre 2011, j'ai étudié l'impact de la polyploïdie sur la divergence d'expression entre les sous-génomes A, B et D du blé hexaploïde à l'échelle du génome entier, à travers le séquençage de son transcriptome.

Cet article, dont je suis le premier auteur, rassemble 5 cosignataires dont il parait important de définir le rôle et la contribution dans les résultats présentés. Ainsi, dans le cadre de cet article, j'ai conduit l'expérimentation en serre pour l'acquisition du matériel végétal nécessaire à l'analyse du transcriptome blé au cours du développement du grain. J'ai mis au point les méthodes d'extraction d'ARN de grain, et d'analyse RNA-SSCP. J'ai réalisé l'analyse transcriptomique en collaboration avec la plateforme Transcriptomique de l'URGV (Sandrine Balzergue). J'ai réalisé l'analyse RNA-seq en prestation de services, auprès de l'entreprise GATC. Enfin, j'ai encadré C. Confolent (CDD de 6 mois) pour la validation des modifications d'expression des copies dupliquées. F. Murat (Bioinformaticien de l'équipe) a réalisé l'analyse de synténie entre le blé et Brachypodium, à partir de laquelle j'ai pu étudier la diploïdisation expressionnelle des copies homéologues chez le blé tendre.

2. Article paru dans la revue 'Genome Biology' en 2011

Pont *et al. Genome Biology* 2011, **12**:R119
http://genomebiology.com/2011/12/12/R119

RESEARCH Open Access

RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.)

Caroline Pont[1†], Florent Murat[1†], Carole Confolent[1], Sandrine Balzergue[2] and Jérôme Salse[1*]

Genome Biology

RESEARCH Open Access

RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.)

Caroline Pont[1†], Florent Murat[1†], Carole Confolent[1], Sandrine Balzergue[2] and Jérôme Salse[1*]

Abstract

Background: Whole genome duplication is a common evolutionary event in plants. Bread wheat (*Triticum aestivum* L.) is a good model to investigate the impact of paleo- and neoduplications on the organization and function of modern plant genomes.

Results: We performed an RNA sequencing-based inference of the grain filling gene network in bread wheat and identified a set of 37,695 non-redundant sequence clusters, which is an unprecedented resolution corresponding to an estimated half of the wheat genome unigene repertoire. Using the *Brachypodium distachyon* genome as a reference for the Triticeae, we classified gene clusters into orthologous, paralogous, and homoeologous relationships. Based on this wheat gene evolutionary classification, older duplicated copies (dating back 50 to 70 million years) exhibit more than 80% gene loss and expression divergence while recent duplicates (dating back 1.5 to 3 million years) show only 54% gene loss and 36 to 49% expression divergence.

Conclusions: We suggest that structural shuffling due to duplicated gene loss is a rapid process, whereas functional shuffling due to neo- and/or subfunctionalization of duplicates is a longer process, and that both shuffling mechanisms drive functional redundancy erosion. We conclude that, as a result of these mechanisms, half the gene duplicates in plants are structurally and functionally altered within 10 million years of evolution, and the diploidization process is completed after 45 to 50 million years following polyploidization.

Background

More than 40 years ago, based on a few protein sequences from vertebrates, Susumu Ohno proposed polyploidization as a major source of new biological pathways created from duplicated gene copies [1]. The vertebrate genomes can be considered as paleopolyploids that had become modern diploids by means of ancestral chromosome fusions as well as sequence divergence between duplicated chromosomes. Recent paleogenomic analyses in plants have confirmed and refined Ohno's conclusions and led to the identification of polyploid common ancestors, showing that present-day species have been shaped through several rounds of whole genome duplications (WGDs), small scale duplications (SSDs) as well as copy number variations (CNVs) of tandem duplicated genes followed by numerous chromosome fusion (CF) events leading to the their present-day chromosome numbers [2-4]. Duplicate genes that persisted in multiple copies diverged by differentiation of sequence and/or function. Overall, recurrent gene or genome duplications generate functional redundancy followed either by pseudogenization (that is, unexpressed or functionless paralogs), concerted evolution (that is, maintained function of paralogs), subfunctionalization (that is, partitioned function of paralogs), or neofunctionalization (that is, novel function of paralogs) during the course of genome evolution. Functional divergence either by subfunctionalization or neofunctionalization of duplicated genes has been proposed as one of the most important sources of evolutionary innovation in living organisms [5]. As a consequence, polyploidy followed by diploidization is a major mechanism that has shaped complex regulatory networks during the

* Correspondence: jsalse@clermont.inra.fr
† Contributed equally
[1]INRA, UMR 1095, Genetics, Diversity and Ecophysiology of Cereals, 234 avenue du Brézet, 63100 Clermont-Ferrand, France
Full list of author information is available at the end of the article

evolution of the plant genomes. However, the real impact of genome duplication on gene network evolution, by comparing ancestral pre-WDG networks to modern post-WGD networks, is not clear. Recent access to numerous sequenced plant genomes [4] now offers the opportunity to study, at an unprecedented resolution, the impact of WGD on gene and genome organization as well as regulation.

Recent paleogenomics studies in plants aiming at comparing modern genome sequences to reconstruct their common founder ancestors based on the characterization of shared duplication events allowed the characterization of seven genome paleoduplications for the monocots and seven genome paleotriplications for the eudicots. These data led to the construction of extinct ancestors of seven protochromosomes (9,731 protogenes) and five protochromosomes (9,138 protogenes) for the eudicots and monocots, respectively [4] (Figure 1a). These recent evolutionary studies in plants suggest that most duplicated genes that are structurally retained during evolution (referred to as 'persistent duplicated genes') have at least partially diverged in their function [6,7]. Microarray studies in eudicots and monocots showed that the vast majority of duplicated genes have diverged in their expression profiles, with 73% [8,9] and 88% [10] of gene pairs in *Arabidopsis* (eudicot reference genome) and rice (monocot reference genome), respectively, associated with asymmetric expression profiles after 50 to 100 million years of evolution. In maize, where a recent WGD dating back to 5 million years ago (MYA) occurred [11], more than 50% of the duplicated genes have been deleted and are no longer detectable within paralogous chromosomal blocks [12]. These results clearly demonstrate that most of the genetic redundancy originating from polyploidy events is erased by a massive loss of duplicated genes by pseudogenization in one of the duplicated segments soon after the polyploidization event.

Because many genes are part of more global regulatory networks, a change in the expression pattern of a single gene could induce changes for numerous genes involved in the same functional pathway. Haberer *et al.* [13] noted for example that tandem as well as segmental duplicate gene pairs exhibiting high *cis*-element similarities within promoters had divergent expression in *Arabidopsis*, suggesting that changes to a small fraction of *cis*-elements could be sufficient for neo- or subfunctionalization. We can argue that functional novelties derived from neo- or subfunctionalization of orthologous and paralogous copies may reduce the risk of extinction of plant species [14,15], similar to what has been suggested in mammals, where extinction events of vertebrate lineages is higher prior to the known ancestral WGD [16]. In this scenario, rapid genomic (that is, reciprocal gene loss) and functional changes (that is, neo- or subfunctionalization) following WGD might enable polyploids to better or quickly adapt to environmental conditions with improved physiological and morphological traits and properties that were not present or sufficient in their diploid progenitors. For instance, it has been suggested that neo- or paleopolyploidy may increase vigor [17], favor tolerance to environmental changes [15], and facilitate propagation through increased self-fertilization species [18,19].

To gain insight into the impact of genome doubling on gene structure and expression, we performed high-throughput RNA sequencing (RNA-seq)-based inference of the grain filling gene network in bread wheat. We focused our functional experiments on a grain developmental kinetic to be able to run comparable experiments in other cereals (for example, rice in the next sections) based on the main conserved grain developmental phases: cell division, filling, and dehydration. Bread wheat is a good plant model to study the impact of distinct rounds of WGD on gene structure and function, as its genome comprises seven ancestral paleoduplications shared with all known cereal genomes and two recent neopolyploidization events to form *Triticum aestivum*, which originated from two hybridizations, one between *Triticum urartu* (A genome) and an *Aegilops speltoides*-related species (B genome) 1.5 to 3 MYA, forming *Triticum turgidum* ssp. *durum*, and one between *T. turgidum* (genomes A-B) and *Aegilops tauschii* (D genome) 10,000 years ago [20,21]. Bread wheat is thus a good genome model to study in the same analysis the impact of ancient and recent WGD on genome structure and function. The bread wheat genome architecture offers us the opportunity to study not only the structures and corresponding expression patterns of paleoduplicated genes (50 to 70 million years of evolution) but also neoduplicated genes (1.5 to 3 million years of evolution) by comparing expression profiles of A, B and D homoeologous gene copies, that is, homoeoalleles (Figure 1a). As the complete assembled wheat genome sequence is not yet available, we have used *Brachypodium* as reference genomes to investigate the grain filling gene network modification in response to recent and ancient evolutionary events, such as duplication, polyploidization and speciation. The aim of this study was not to perform a quantitative (that is, transcriptome) analysis of the genes expressed during grain development but rather a robust qualitative identification (that is, large scale repertoire) of homoeologous/orthologous/paralogous gene networks, allowing us to provide new insights into the structural and functional evolution of genes after a WGD event in plants. This article provides relevant conclusions on how recent and ancient duplicated genes in plants evolve in both

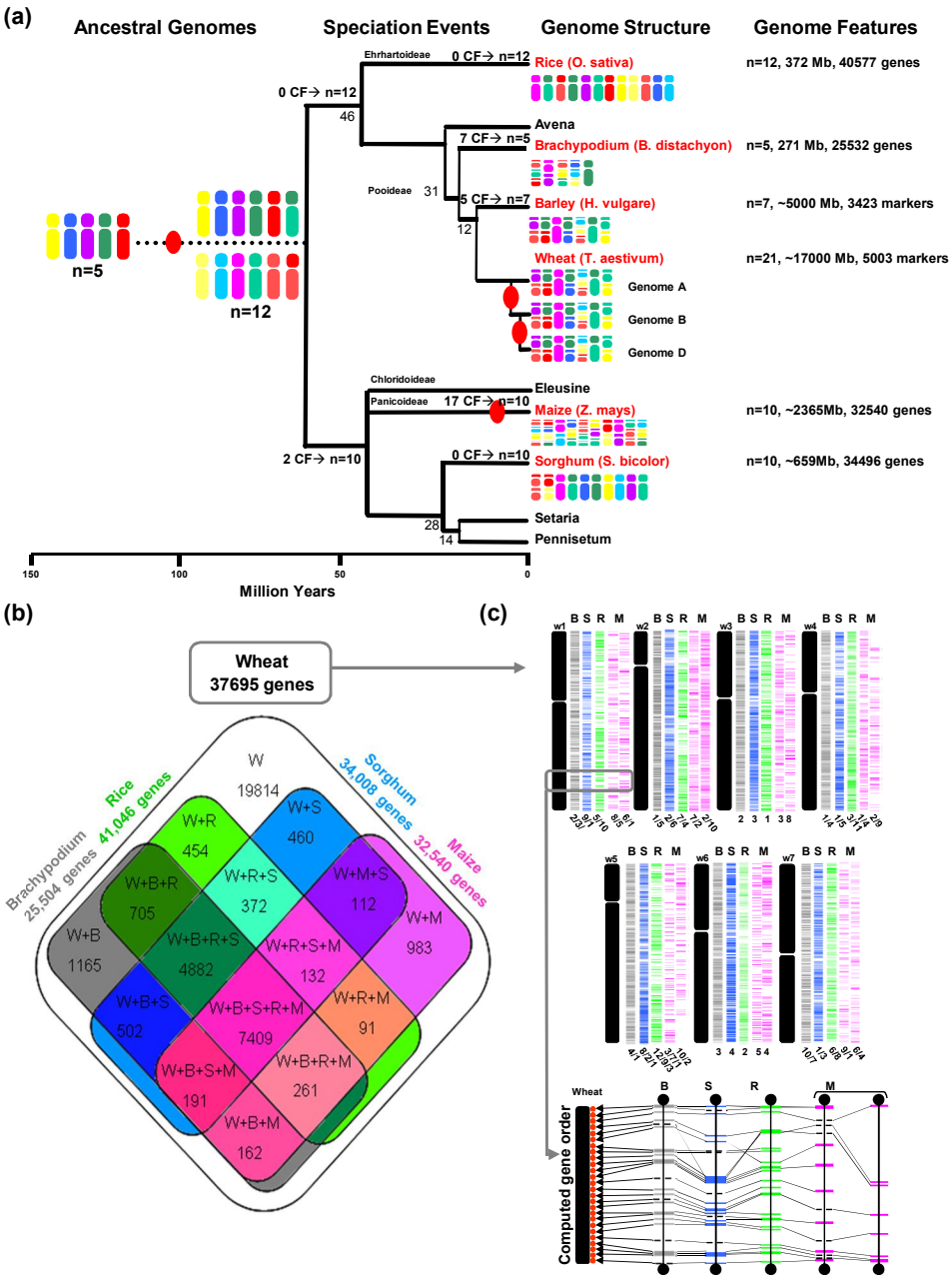


Figure 1 Homolog gene conservation between wheat and cereal sequenced genomes. **(a)** Cereal genome paleohistory. Schematic representation of the phylogenetic relationships between grass species adapted from [2,4]. Divergence times from a common ancestor are indicated below the branches of the phylogenetic tree (in million years). Whole genome duplication events are illustrated with red circles on the tree branches. The evolution of chromosome numbers of modern species from the ancestral genome structure is indicated with the number of chromosome fusion (CF) events. Genome features (number of chromosomes, physical size, and the number of annotated unigenes) of the six cereal genomes investigated are shown at the right-hand side. Modern genome architectures are illustrated using a color code that represents the $n = 5$ and 12 extinct ancestors (left). **(b)** Homologous gene groups between wheat and rice, *Brachypodium*, sorghum, and maize genomes. The Venn diagram illustrates the number of conserved protein domain-based homologs between wheat (RNA-seq gene clusters) and rice/*Brachypodium*/sorghum/maize (annotated proteins). **(c)** Simulated synteny-based gene order model in bread wheat. The chromosomal location of the RNA-seq gene clusters are shown on the seven bread wheat chromosome groups based on a consensus gene order derived from the observed synteny between wheat and rice ('R', in green), *Brachypodium* ('B', in grey), sorghum ('S', in blue), and maize ('M', in pink) chromosomes (numbers are shown at the bottom of the chromosomes). The bottom inset illustrates a micro-synteny example of 26 re-ordered genes in bread wheat chromosome 1 (red dots) based on orthologous genes identified in *Brachypodium* (chromosome 2, 92 annotated genes, 0.9 Mb), sorghum (chromosome 9, 108 annotated genes, 1.1 Mb), rice (chromosome 5, 112 annotated genes, 0.9 Mb), maize (chromosomes 6 to 8, 145 annotated genes, 12.6 Mb). Non-conserved genes are illustrated using dotted lines and conserved genes are linked with black lines.

structure and function at the whole genome level, the gene family level, and the gene network level. The established divergence of structural and expression patterns between duplicated genes might have accelerated the erosion of colinearity between plant genomes as discussed in the article.

Results

Synteny-based gene repertoire and expression map in wheat

We performed an RNA-seq analysis of samples collected during the grain development in wheat. We used a 454 (Roche, see Materials and methods)sequencing platform with five developmental stages, that is, 100 degree days (DD), 200 DD, 250 DD, 300 DD, and 500 DD after pollination. The five developmental stages cover the cell division (100, 200, 250 DD) and filling (300, 500 DD) phases of grain development in wheat. RNA was extracted, pooled, and sequenced and sequence reads (934,928 in total) were clustered and checked for quality as described in the Materials and methods section in order to provide a qualitative and exhaustive view of the grain development gene network in bread wheat (Table S1 in Additional file 1). We obtained 37,695 sequence clusters (20.1 Mb of assembled sequences with an average coverage of approximately 25× per cluster) based on the assembly strategy protocol described in the Materials and methods section. Detailed information on the 37,695 sequence clusters (identity, sequence, and function) is available in Table S2 in Additional file 1 and consists of the most complete gene network repertoire of the grain development in wheat and probably in grasses more generally.

We aligned the 37,695 sequence clusters to the proteomes of the four monocot sequenced genomes, that is, rice, sorghum, *Brachypodium*, and maize (Figure 1b). Homologous gene pairs based on protein sequence conservation (BLASTx) of functional domains allowed us to establish that 17,881 (47%) wheat genes can be paired with a single homolog counterpart (based on sequence comparisons using 50% protein identity as a threshold criterion) in at least one of the considered sequenced genomes. The remaining 19,814 are putative wheat-specific unigenes (that is, not found in any of the four sequenced cereal genomes available to date) based on our BLAST alignment criteria, including 8,428 (43%) associated with wheat public EST-unigenes and 11,386 short reads (that is, an average of 430 bases for wheat-specific versus 650 bases for non-wheat-specific clusters) and/or low expressed/covered genes (that is, an average of 15× for wheat-specific versus 36× for non-wheat-specific clusers). We cannot finally exclude that such orphan clusters may correspond to sequenced poly-adenylated non nuclear sequences. As expected, the *Brachypodium* sequence genome appears to be the closest relative with the highest number of specific (not shared with any of the three other sequenced cereal genomes) protein-based homologs (1,165) identified in comparison with the wheat unigene set. A four genome-based synteny approach was used for all seven wheat chromosome groups by integrating wheat cytogenetic map information [22] and public chromosome-to-chromosome relationships [2,4] to produce the most parsimonious simulated gene order in wheat based on gene conservation observed among the four sequenced cereal genomes as detailed in Murat *et al.* [23]. Based on the known synteny relationship established between the seven wheat chromosome groups and the rice, sorghum, *Brachypodium* and maize genomes [23], we produced a partial wheat gene-based physical map where RNA-seq clusters were ordered within wheat chromosomes in respect to the position of their orthologous counterparts (following the ordering priority of rice >*Brachypodium* > sorghum > maize; Figure 1b; Table S3 in Additional file 1). A comparable approach has also been used recently in barley [24]. The gene content for chromosome 3B has recently been estimated to include 8,400 unigenes [25], of which 3,478 (41.4%) were available from the current analysis. We provide here the largest set of unigenes in wheat, covering almost half of the total genome-wide gene set based on the previous 3B chromosome comparison. Our wheat unigene set originated from a single tissue (grain), suggesting that only a few additional complementary ones (such as from root and leaf) would be sufficient to recover the vast majority of all genes in wheat. Therefore, we were able to place 17,881 wheat genes in a so-called computed or simulated order along chromosomes (Figure 1b) and have made the data available to users (Table S3 in Additional file 1) for further marker development or candidate gene identification. Figure 1c (bottom inset) illustrates the strategy used to infer computed gene order in wheat (chromosome group 1) based on the consensus gene order derived from the synteny observed between *Brachypodium* (chromosome 2), rice (chromosome 5), sorghum (chromosome 9) and maize (chromosomes 6 to 8) genomes. We therefore provide here for the first time the most complete qualitative set of unigene sequences expressed during the grain development in wheat associated with synteny-based physical locations on the seven chromosome groups.

Using the *Brachypodium* genome (5 chromosomes, 271 Mb, 25,504 gene models) as a reference to produce a heterologous wheat expression map, we could identify one-to-one robust orthologous gene pairs between wheat RNA-seq clusters and *Brachypodium* gene models using two nucleic acid alignment (BLASTn) parameters as described in Salse *et al.* [2,3] and the Materials and

methods section. Briefly, with the BLASTn alignment based on default parameter (such as expect values), homologous gene relationships are obtained, although the analysis is polluted with background noise corresponding to high functional domain conservation, making it difficult to characterize which are the real significant single orthologous relationships between two considered genomes. The used parameters (CIP = 60% and CALP = 70%) return statistically significant single copy collinear relationships between two gene sets, and the remaining homologous gene relationships are then considered artifactual, that is, obtained at random [2]. Among the 37,695 RNA-seq clusters, 8,485 (23%, with an average size of 761 bases) wheat sequences could be aligned with 7,158 known orthologous genes in *Brachypodium* (Table S4 in Additional file 1) following this strategy. Map positions in wheat were simulated from syntenic relationships with *Brachypodium* as explained in the previous section. The remaining 29,210 RNA-seq clusters that could not be paired with *Brachypodium* gene models corresponded to short reads (average size of 468 bases) that were either considered as singletons or rejected based on our stringent sequence alignment criteria. These stringent alignment criteria were set to establish a robust repertoire of homoeologous/paralogous (wheat), orthologous (wheat/*Brachypodium*) genes in order to infer the consequence of evolutionary events (duplication, speciation) on gene structure and expression patterns, as discussed in the next sections. The objective of the current analysis was not to obtain the largest set of wheat homologous counterparts in *Brachypodium* for the 37,695 wheat sequence clusters (as described in the previous section and illustrated in Figure 1b) but rather precise and robust evolutionary relationships (conserved and duplicated genes) to investigate structural and functional redundancy.

In summary, we produced 37,695 wheat gene clusters (estimated to represent half of the total diploid wheat gene content based on the wheat chromosome 3B-based inference), of which 47% were associated with functional domain-based homologs of the sequenced genome proteomes (*Brachypodium*, rice, sorghum and maize) and 23% were strict orthologs with *Brachypodium*, considered as the sequenced reference genome for the Triticeae.

Evolutionary fate of duplicated genes at the whole genome level

We produced a heterologous wheat expression map where 8,485 genes that were expressed during wheat grain filling were mapped strictly to the *Brachypodium* genome and positioned within the wheat genome based on the recently established *Brachypodium* /wheat genomes colinearity [4] (Figure 2a). This heterologous wheat expression map (Table S4 in Additional file 1) has been used to study and discuss the evolutionary fate of paralogous, homoeologous and orthologous gene copies. Figure 2a depicts the five *Brachypodium* chromosomes as the inner circle (labeled 'Bd') and illustrates the seven paleoduplications (in black) shared with other cereals in the center [2,3]. The second circle (labeled 'Ta') illustrates the orthologous relationships identified between *Brachypodium* and wheat using a seven color code [4], illuminating the Triticeae chromosome group origins. Black dots around the wheat circle illustrate the 454 RNA-seq reads from wheat (labeled '$454_{Ta'}$).

Our data clearly show that 6,024, 941, and 193 gene models matched with 1, 2, and 3 homoeologous gene copies in wheat, respectively (Figure 2a, 454Ta circles). Overall, only 193 of 7,158 orthologous gene pairs identified between *Brachypodium* and wheat matched the three expected homoeologous counterparts in wheat. Therefore, we can suggest that 2.7% of the homoeologous copies derived from two rounds of polyploidization that took place less that 1.5 to 3 MYA [20,21] have been structurally and functionally conserved in wheat. We cannot exclude that the expression of some homoeologs may be too low to be detected by RNA-seq given the coverage used in the current analysis, but we have clearly established that, for 6,024 genes with detectable expression signals, the three homoeologs do not have perfectly redundant expression profiles in the considered grain experiment. This clearly suggests that, for a large majority of the homoeologs in wheat, at least one copy has been lost (deleted or pseudogenized) or neo- and/or subfunctionalized within 1.5 to 3 million years of evolution. Moreover, Figure 2a illustrates that the genes expressed during the grain development were randomly distributed on the *Brachypodium* genome. Because of the average RNA-seq cluster length, however, we cannot distinguish homoeologous copies that have SNPs (1 per approximately 500 base-pairs in wheat) outside of the aligned sequences, leading to an overestimated percentage of homoeologous gene rearrangement events through homoeologous gene assemblage in the same cluster. We can still hypothesize that such homoeologs that do not harbor homoeoSNPs within the sequenced regions have been clustered together, leading to increased sequence coverage of the considered clusters. Consequently, we can extrapolate the homoeologous representation within a sequence assembly cluster based on the sequence coverage. The assumption is that for a sequence fraction of a gene that does not harbor homoeoSNPs, the three homologs are then clustered within the same assembly, leading to an increase in the sequence coverage of such a region. Monitoring putative merged homoeologs in the same clusters based on the sequence coverage of the initial 37,695 RNA-seq

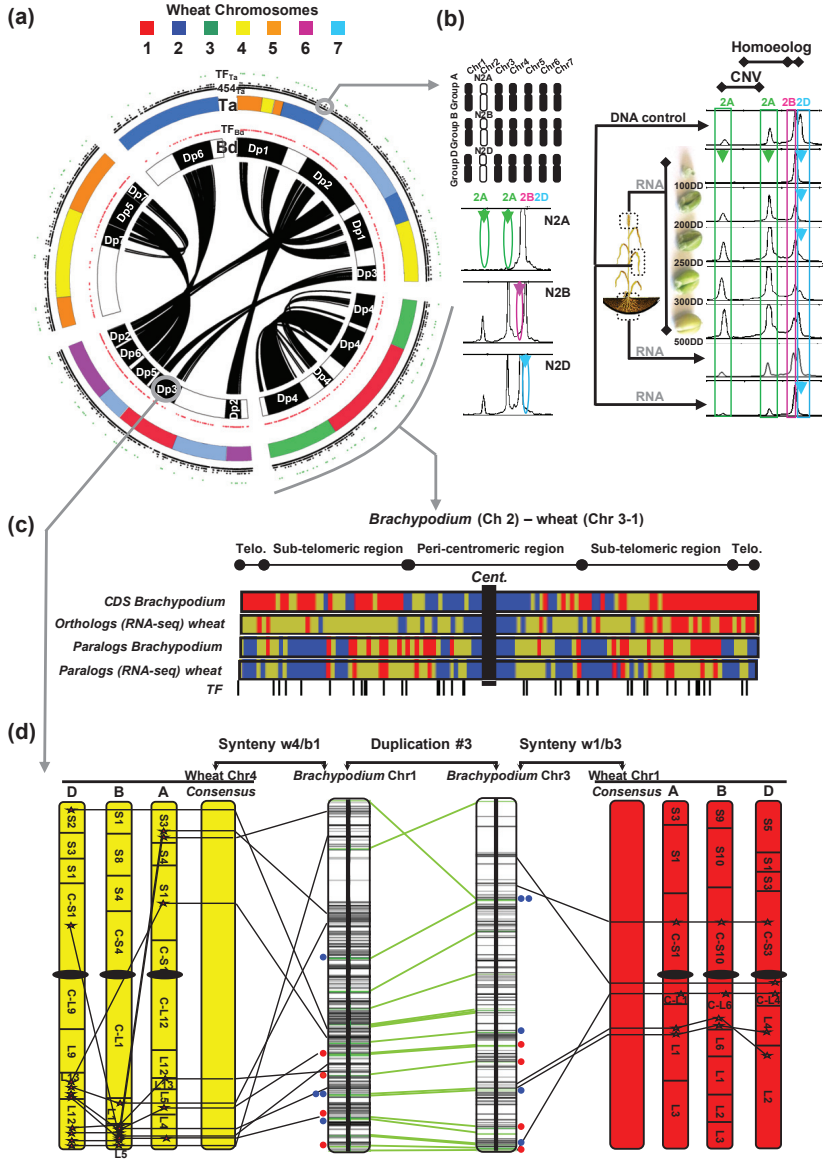


Figure 2 Heterologous genome-wide wheat expression map. **(a)** The five *Brachypodium* chromosomes are illustrated in the inner circle (labeled Bd) and the seven paleoduplications (in black) shared within cereals are displayed in the center (Dp1 to Dp7). The second circle (labeled Ta) illustrates the orthologous relationships identified between *Brachypodium* and wheat using a seven color code. Dots around the two circles illustrate the *Brachypodium* transcription factors (red; labeled TF_{Ta}), the 454 RNA-seq reads from bread wheat (black; labeled 454_{Ta}) and associated TFs (green; labeled TF_{Bb}). **(b)** CNVs and homoeologous gene localization in bread wheat is illustrated with a single COS marker (CT753726) that has been located on chromosome 2A (CNV of 2), 2B and 2D using the adapted cytogenetic material illustrated in the top. The arrows illustrate the observed amplification loss observed for the illustrated cytogenetic material (N2A, N2B, N2D respectively for the absence of the 2A, 2B, 2D chromosomes). The COS marker (CT753726) expression (SSCP amplification) profiles observed in the five RNA samples from wheat grain development, as well as in leaves (RNA and DNA amplifications) and roots (RNA) considered as negative control. Colored arrows highlight the loss of expression of the considered CNV or homoeologous copies. **(c)** Wheat (chromsomes 1 to 3) and *Brachypodium* (chromosome 2) heat maps for *Brachydopium* coding sequence (CDS; blue < 40, yellow approximately 41 to 50, red > 51 genes/500 kb), wheat RNA-seq ortholog (blue < 9, yellow approximately 10 to 19, red > 20 genes/500 kb), *Brachypodium* paralog (blue = 0, yellow approximately 1 to 5, red > 6 genes/500 kb), wheat RNA-seq paralog (blue = 0, yellow approximately 1 to 3, red > 4 genes/500 kb) distributions. The 44 RNA-seq TFs are illustrated with their corresponding orthologous positions on the *Brachypodium* chromosome as black vertical bars. **(d)** Paralogous chromosomal regions are shown in the center, involving *Brachypodium* chromosome 1 (2.1 Mb, 252 genes) and 3 (2.9 Mb, 181 genes), and annotated genes are shown with horizontal bars. Orthologous wheat chromosomes are shown at the right (consensus group 1 and homoeologous chromosomes in red) and left (consensus group 4 and homoeologous chromosomes in yellow). Orthologous genes identified between wheat and *Brachypodium* are linked with black lines. Paralogous genes identified between *Brachypodium* chromosomes are linked with green lines. Expression data from wheat RNA-seq cluster alignment against the considered *Brachypodium* chromosome sequences are shown with colored dots. Blue dots illustrate paralogous gene pairs for which only one copy is associated with wheat RNA-seq clusters while red dots illustrate paralogous pairs for which both duplicates are associated with wheat RNA-seq clusters.

unigenes, we were then able to identify 1,009 clusters covered with more than 140 reads (that is, putatively merged homoeologs) compared to 7,158 single copy homologs reported previously and associated with an average coverage of 38 to 42 reads. We confirm at the whole gene repertoire level (37,695 non-redundant sequence clusters), following a sequence coverage-based approach for detecting homoeologous copies, what we observed initially using a synteny-based approach (2.7% of wheat homoeologs are associated with a *Brachypodium* orthologous gene), that 2.7% (1,009) of the homoeologous copies derived from two rounds of polyploidization that took place less than 1.5 to 3 MYA have been structurally and functionally conserved.

We designed an experiment using a subset of wheat genes to confirm the *in silico* structural and functional inference of homoeologous gene copies based on an *in omic* complementary approach. It is possible that missing data (non-sequenced low expression genes) have led to an overestimation of the structural and expression differences between homoeoalleles. If we select a subset of 100 genes, we should observe, based on the *in silico* conclusions detailed in the previous sections, that a vast majority of the homoeologs do not share the same expression pattern during grain development. To do so, from a set of 100 wheat genes randomly distributed among the 7 chromosome groups, we were able to design 91 primer pairs for further *in omic* structural (that is, evidence of homoeolog deletion) and functional (that is, evidence of homoeolog neo- and subfunctionnalization) inference of homoeologous gene copies in bread wheat. The Single Strand Conformational Polymorphism (SSCP) detection allows identification of homoeologous amplicons in a polyploidy background through the exploitation, on a capillary sequencer, of secondary DNA structure under non-denaturing conditions. The SSCP approach on a capillary sequencer [26] offers two advantages, the ability to detect SNP and size polymorphisms and to identify homoeologous or even paralogous amplifications. Using the wheat cytogenetic material available for the structural detection of putative homoeologs based on the SSCP technique, we observed that 43 (54%) out of 79 successfully assigned genes exhibited loss of at least one homoeologous copy beyond technical detection. Regarding expression patterns, 33 genes (36%) out of 91 showed a loss of expression when considering grain development, whereas 45 (49%) showed a loss of expression at the whole tissue level when comparing the expression in grain, the leaf, and the root (Figure 2b; Table S5 in Additional file 1). Consequently, 49% of the wheat homoeologous gene copies have been neo- and/or subfunctionalized when considering the grain developmental kinetic. Finally, only 27 (34%) genes out of 79 homoeoalleles detected on the three chromosome groups clearly show a conserved expression pattern in grain. The remaining 66% have either been structurally lost and/or neo- and/or subfunctionalized in their expression profiles. Figure 2b (left) illustrates the chromosomal localization of a single COS (conserved orthologous set [26]) gene (wheat CT753726 with rice ortholog LOC-OS04g33150) assigned to chromosome group 2 (homoeologous copies A, B and D as well as a single CNV for the A homoeolog). The same COS gene used to amplify, through the SSCP approach, the five RNA samples clearly shows that either the homoeologs (A, B and D copies) or CNVs do not present a perfect redundancy in their expression patterns. Figure 2b (right) illustrates how homoeoalleles and CNV expression signals were alternatively lost during grain development (colored arrows). Therefore, if 66% of homoeologs in wheat were either structurally lost (54%) or have diverged in their expression patterns (36% within tissues and 49% between tissues), earlier *in silico* assessments of homoeologous gene shuffling (that is, only 2.4% of homoeoalleles show conserved expression profiles) deduced from the alignment-based construction of homoeologous RNA-seq clusters was indeed overestimated (by about 20 to 30%), probably because of an average sequence read length of 761 bases as well as the possibility of missing low expressed genes, limits associated with this sequencing strategy.

Figure 1a illustrates a non-random distribution of wheat/*Brachypodium* orthologous genes at the whole genome level. As an example, Figure 2c shows *Brachypodium* chromosome 2, where the first heat map (coding sequence ('CDS') track) illustrates the distribution of annotated CDS with a clear enrichment of CDS in subtelomeric regions (that is, 107.2 genes/Mb) and a reduced density in peri-centromeric regions (that is, 65.3 genes/Mb) due to transposable element (TE) invasion [27]. The second heat map illustrates the density of *Brachypodium* genes associated with a wheat ortholog ('Orthologs' track) based on the data set of 8,485 RNA-seq clusters (Table S6 in Additional file 1). The gene conservation is higher in peri-centromeric regions (31.1% of conserved genes) compared to telomeric (23.8% of conserved genes) or sub-telomeric (28.1% of conserved genes) regions. Finally, the paralogs (either *Brachypodium* or wheat gene 'paralogs' tracks) are not randomly and homogeneously distributed among chromosomes, that is, 47.4% versus 79.2% of duplicated genes in telomeric versus sub-telomeric regions, respectively. The 862 duplicated genes in *Brachypodium*, which arose from the seven ancestral duplications shared by the Poaceae, are depicted in the center of Figure 2a (from 1 to 7). Therefore, 166 *Brachypodium* paralogous pairs (19.3%) matched with their duplicated

counterpart in wheat. The remaining 696 paralogous pairs (80.7%) matched with no or only one wheat sequence derived from the RNA-seq repertoire. This result is consistent with previous results [10] showing that 87.4% of the paleoduplicated genes in rice have been lost within a 50 to 70 million years of evolution. Figure 2d provides a detailed view of the ancestral duplication referenced as 'Dp3' shared between *Brachypodium* chromosomes 1 to 3 and wheat chromosomes 1 to 4. Duplicated genes are connected with a green line at the center of Figure 2d and wheat RNA-seq clusters that are orthologs of *Brachypodium* duplicated genes are illustrated with blue dots (wheat homoeologous genes identified for only one of the *Brachypodium* duplicates) or red dots (wheat homoeologous genes identified for both of the *Brachypodium* duplicates). At a micro-scale level for one (Dp3) of the seven ancestral duplications, among 20 paralogous gene pairs (green lines), 4 (20%) matched wheat homoeologous gene copies expressed during grain development (Figure 2d, red dots). This result further refines the conclusion that at either the whole genome level (19.3% of duplicates with concerted expression in the grain) or the micro-scale level (20% of duplicates with concerted expression in the grain), most of the paleoduplicated genes have been either lost or neo- and/or subfunctionalized so that the expression patterns at the tissue level are no longer redundant.

In summary, despite limitations of the RNA-seq approach in detecting low expressed genes and differentiating homoeoalleles, we have clearly shown at the whole genome level, using a heterologous wheat expression map, that almost 70% of recent duplicates (from homoeologous copy evolutionary analysis) have diverged during 1.5 to 3 million years of evolution (54% of homoeologous copies structurally lost and 36 to 49% of homoeologous copies with different expression profiles), and that more than 80% of ancient duplicates (from paralogous evolutionary analysis) have diverged during 50 to 70 million years of evolution.

Evolutionary fate of duplicated genes at the gene family level

Out of the 7,158 *Brachypodium* genes corresponding to 1, 2, or 3 wheat homoeologous gene copies derived from the grain RNA-seq data described previously, 5,967 (corresponding to 7,112 wheat sequences) follow a canonical Gene Ontology (GO) classification. Among the 38 GO categories (from 'molecular function' classification) described at the whole genome level in *Brachypodium*, the distribution of three classes were shown to be statistically (based on chi-square test using 1% as a threshold) biased between grain development data (that is, from wheat RNA-seq) and what is observed at the whole-genome level (that is, annotated genes in *Brachypodium*): protein binding, transcription factor activity, and electron carrier activity (Table S7 in Additional file 1). The three previous GO classes are then good candidates to study the evolutionary fate of duplicated genes at the gene family level.

We recently performed a transcriptome analysis of rice grain filling based on an oligonucleotide array, where among the 60,727 genes spotted on the array, 29,191 were expressed during grain development [10]. In particular, we conducted a detailed analysis of 32 transcription factors (TFs) that were expressed during rice grain development. Across 100 to 600 gene physical intervals covering the entire rice genome, no co-regulation was observed between the selected TFs and the flanking genes [10]. In order to test this hypothesis, we conducted a specific analysis of TF gene families in wheat. Among the 666 TFs identified in the *Brachypodium* genome annotation (Figure 2a, TF_{Bb} red dots), 161 wheat homoeologs were extracted from the RNA-seq clusters (Figure 2A, TF_{Ta} green dots). Of these 666 *Brachypodium* TFs, 140 (21%) matched with a wheat ortholog that was expressed during grain development. Figure 2c shows a classical heat map representation of *Brachypodium* chromosome 2, including the distribution of 44 TFs (Table S8 in Additional file 1) from the wheat RNA-seq clusters that matched an orthologous counterpart of *Brachypodium* chromosome 2 (highlighted with black bars). As can be observed, whereas the distribution of genes among *Brachypodium* chromosomes is concentrated in the subtelomeric regions (see the previous section), the TFs are conserved in orthologous positions along the entire chromosome (that is, 0.5 TF/Mb in telomeric regions versus 0.6 TF/Mb in centromeric regions). These data complement and refine earlier conclusions about wheat diploidization-resistant genes, that is, genes that are preferentially conserved among cereal after WGD are TFs or TF-related gene functions [3,28], leading to a random distribution of this gene family in modern genomes.

In summary, we have established that, at the gene family level and using the TF family as a reference, that the GO 'transcription factor activity' class could be considered a diploidization-resistant gene function as it might have provided a selective advantage during evolution and adaptation and then retained as functional after WGD. Our data support preferential structural conservation of duplicated genes involved in signal transduction and more precisely transcription, which are putatively involved in response to rapidly changing biotic and abiotic extrinsic factors compared with genes encoding products involved in relatively more stable processes.

Evolutionary fate of duplicated genes at the gene network level

In order to compare the wheat grain filling gene network with the previously published rice transcriptome analysis [10] described in the previous section, we conducted a similar analysis using the wheat Affymetrix array (based on the design and methods described in Wan *et al.* [29]). To avoid any bias due to different expression analysis methods - that is, RNA-seq versus Array technologies - we used the same RNA samples from the five wheat grain stages for hybridization of the wheat Affymetrix array (based on two independent biological replicates), see Materials and methods. Among the 6,760 rice/wheat transcripts identified between the rice (60,726 oligonucleotide probes) and wheat (61,115 oligonucleotide probes) arrays, 2,600 (38.4%) showed concerted (that is, Presence versus Absence Variation, referenced as PAVs) expression signals during grain development (Table S9 in Additional file 1). When considering not only rice/wheat orthologs but also paralogs that might have conserved the original or ancestral gene function and expression in rice and wheat, the percentage of concerted expression between both species would increase to 43.5% (that is, 2,944 genes).

Among the plant metabolic networks, the starch synthesis pathway is well known because starch is considered a major key regulator of grain development. In this network, 170 enzyme-coding genes can be represented with nodes and substrate-product metabolite flux by directional edges [30]. Figure 3 illustrates the comparative gene network observed between rice (gene profiles from microarray data [10]) and wheat (expression data from the current oligo-array and RNA-seq data) for the starch biosynthesis pathway described in Zhu *et al.* [31] (Table S10 in Additional file 1). Among the 170 genes involved in this network, 24 (14%) were identified as differentially expressed in rice and wheat based on the microarray experiments. However, based on the wheat RNA-seq data, 84 (49%) of the 170 enzyme-coding genes could be matched with 1 (57 genes), 2 (21 genes) or even 3 (6 genes) homoeologous copies. We also could show that among the 84 genes for which we have identified RNA-seq clusters as proof of expression in grain development, only 6 (7%) matched their three homoeologous counterparts. This micro-scale analysis, focused on a unique and specific well-known gene network, also agreed with the whole-genome level analysis that revealed that, for a large majority of the homoeologs in wheat, at least one copy had been lost or neo- and/or subfunctionalized during 1.5 to 3 million years of evolution.

Figure 3b illustrates the impact of the paleoduplication in grasses on the starch network. Based on the identification of 20 duplicated genes (black brackets in Figure 3a, 'Gene Acc' column) within the 170 enzyme-coding genes, we can suggest that 12% of the actual modern post-WGD network has been enriched by the ancestral shared tetraploidization event. We can then model the ancestral pre-WGD network consisting of 150 non-redundant starch enzyme-coding genes (Figure 3b). The observation that the post-WGD network is more abundant and enriched in TFs is also consistent with previously reported biases in gene functions after WGDs in plants [32,33] as well as in fungi and mammals [7,34]. Previous results for cereal genomes [2-4,28] and for eudicots [32] clearly showed that retained duplicated gene families correspond to transcriptional regulators that were preferentially conserved after WGD events. However, our analysis, based on a single gene network, did not confirm earlier reported conclusions in *Arabidopsis* that bottleneck enzymes in metabolic networks, which tend to connect different modules, are preferentially retained as functional duplicates after WGD [35]. In our case, of the seven genes preferentially retained as duplicated (highlighted in red in the post-publication network representation), none correspond to enzyme-node encoding genes.

In summary, we suggest that, at the gene network level and using the starch biosynthesis pathway as a reference, 14% of the rice-wheat orthologous copies have the same expression pattern (compared to up to 44% at the whole-genome level), 7% of the wheat homoeologous triplicates share the same expression pattern (consistent with what is observed at the whole-genome level), and WGDs have enriched the starch gene network by up to 12% in gene content.

Evolutionary consequences of duplicates on genome colinearity

Structural rearrangement and gene loss between duplicated regions results in the reduction of orthologous relationships between cereal genomes. Duplicated gene loss in maize (Figure 1b, bottom inset) accounts for the major source of erosion of colinearity between maize and the other grass genomes. Gene colinearity observed between maize chromosome 8 (or 6) is reduced compared to the microsynteny observed between *Brachypodium*, rice, and sorghum at the same loci due to the recent WGD that occurred specifically during maize genome evolution. Only seven chromosome 8 (purple) and eight chromosome 6 (purple) genes are conserved between maize and the other three cereal genomes compared to the 26 orthologous relationships (grey, blue, green in Figure 1b) identified when comparing the rice, sorghum, and *Brachypodium* genomes.

Despite the diploidization process following WGD associated with the loss of homoeolog and/or paralog sister gene copies being the major source of genome

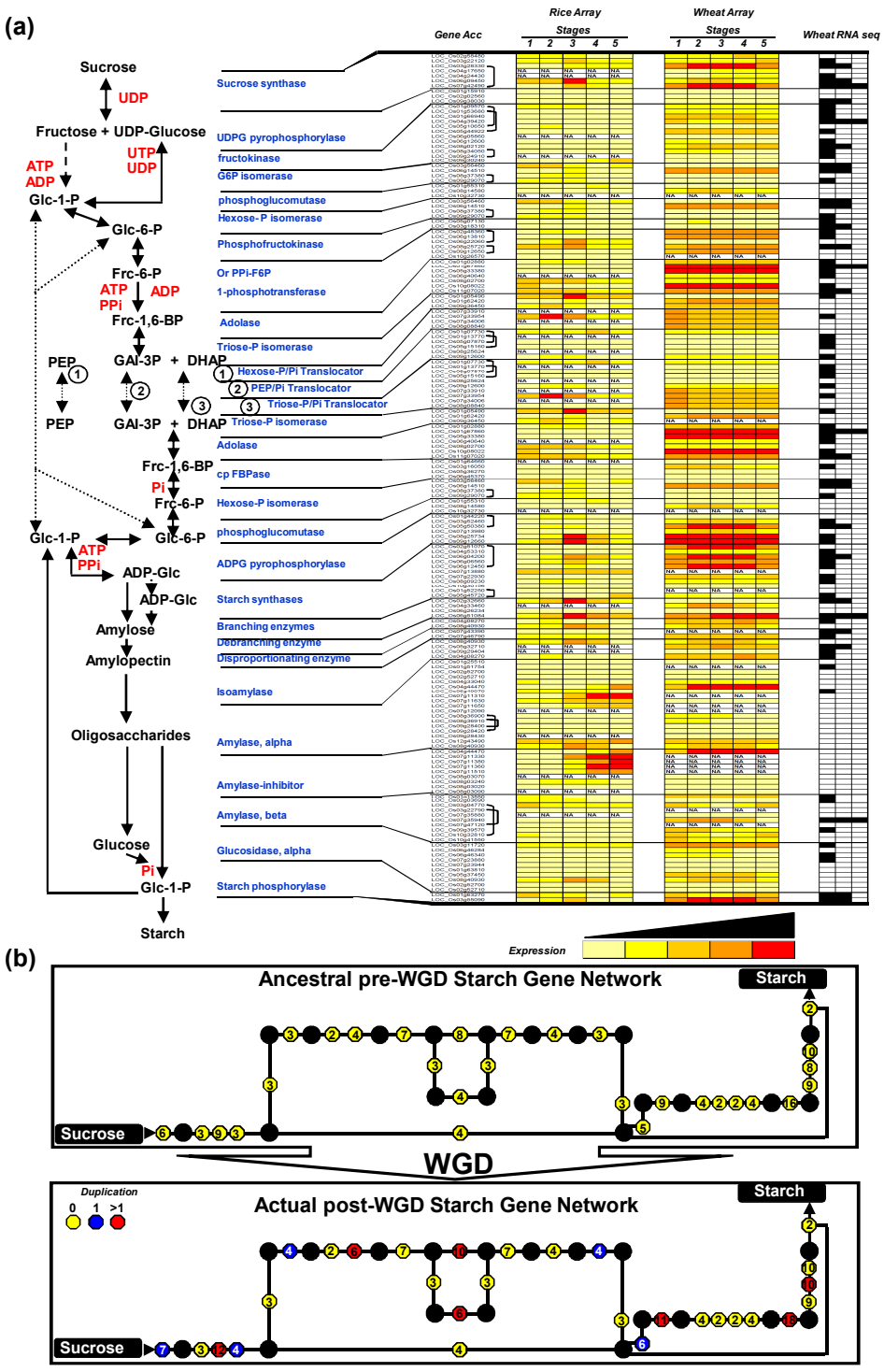


Figure 3 Comparative evolution of the starch gene network between rice and wheat. **(a)** Twenty-nine enzyme-encoding genes involved in the starch biosynthesis network are listed at the left and 170 genes (named according to their accession numbers) are shown. Expression data from oligo-array experiments performed in rice and wheat are displayed using a classical colored scale in which yellow illustrates the absence of expression and red the observed highest level of expression (6.5 for rice, 14.5 for wheat as normalized intensity). At the right are illustrated orthologous RNA-seq clusters from one up to three homoeologous copies (black scares). **(b)** The starch gene network is schematically represented with black dots representing synthesized substrate and the number of gene-encoding enzymes is shown in circles. The pre-WGD (top) and post-WGD (bottom) networks illustrate the impact of paleoduplication where no impact (yellow), gain of an additional enzyme-encoding gene (blue) and more than one additional enzyme-encoding gene (red) are shown.

colinearity erosion, CNV is also an important phenomenon that can contribute to the observed reduced synteny between grass genomes, as illustrated in the Figure 1b where a single-copy COS gene identified in rice, *Brachypodium*, maize and sorghum corresponds to a putative CNV in bread wheat chromosome 2A. Such species-specific CNVs will not be associated with any orthologous counterpart in the other genomes, thus reducing the percentage of conserved and orthologous genes in grasses. Figure 4 illustrates the difference between the loss of synteny and the increased number of tandem duplications, which were referred to as CNVs. At the right-hand side of the figure, bread wheat chromosome 3B is shown with 8 deletion bins, for which the number of available ESTs (blue bars) as well as the number of orthologous genes (red bars) with *Brachypodium* chromosome 2 is illustrated. The orthologous blocks observed between wheat chromosome 3B and *Brachypodium* chromosome 2 are illustrated in different colors in the center of the figure. Finally, at the left-hand side of the figure, *Brachypodium* chromosome 2 is split into orthologous blocks of bread wheat chromosome 3B. The number of RNA-seq clusters of *Brachypodium* genes is depicted as circles (black for homoeologous copies and red for CNVs). A clear correlation between the loss of colinearity and increase of CNV can be observed. The three 3B bin intervals displaying the highest loss of colinearity (3BS8, 3BL2, 3BL7; indicated by red stars) are associated with orthologous regions of *Brachypodium* chromosome 2 comprising CNVs (linked with dotted black lines). Considering *Brachypodium* chromosome 2 as an example of a reference and model chromosomal structure, CNVs in wheat were preferentially located within subtelomeric regions of modern chromosomes or paleo-inserted chromosomes (that is, the ancestral fusion event between W3 in green and W1 in red). We suggest here that the loss of colinearity observed locally between

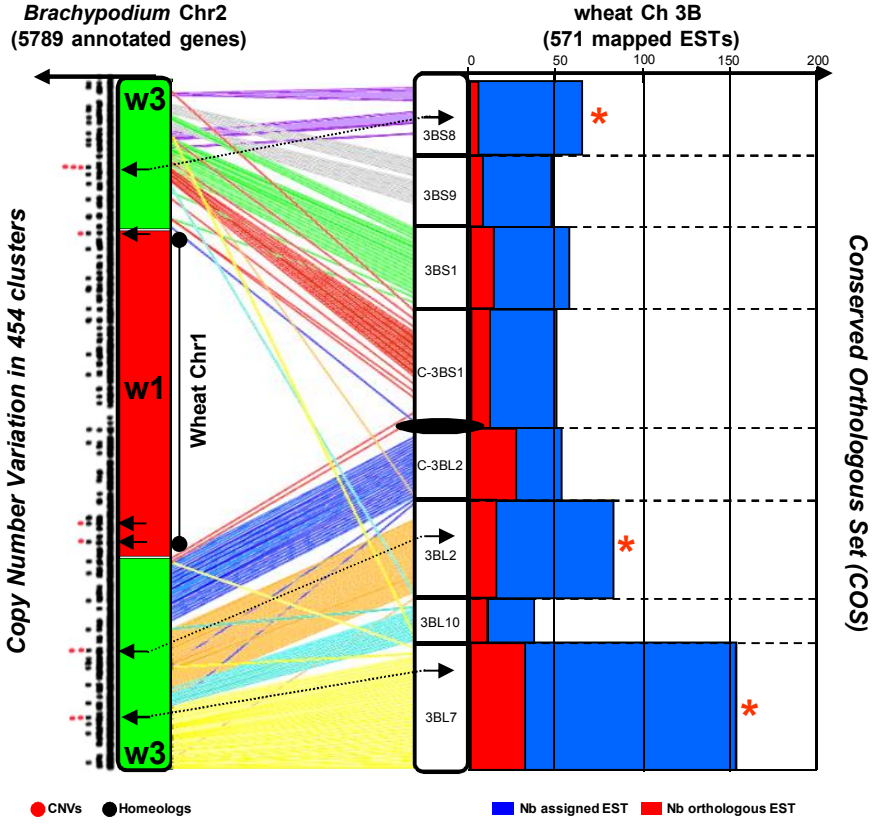


Figure 4 Correlation between CNV and synteny erosion between *Brachypodium* and wheat. At the right is illustrated the wheat chromosome 3B subdivided into eight deletion bins surrounding the centromere. The number of available mapped ESTs per bin is illustrated with blue bars whereas the number of conserved orthologous sets (COS) is shown with red bars. The three deletion bins associated with the lowest percentage of COS regarding the number of available mapped ESTs are indicated by red stars. At the left is illustrated the *Brachypodium* chromosome 2 subdivided into orthologous chromosomal relationships with wheat chromosomes 1 (in red) and 3 (in green). The wheat RNA-seq clusters are shown according to their orthologous positions on *Brachypodium* chromosome 2 as homoeologs (black dots) or additional CNVs (red dots). At the center are shown the *Brachypodium*/wheat orthologous genes linked with colored lines according to their position within the wheat deletion bins.

Brachypodium and wheat is mainly due to tandem gene duplications putatively favored by recent polyploidization events in bread wheat.

In summary, we have shown that, at the whole genome as well as the chromosome level, segmental duplications and gene duplications in tandem (CNVs) comprise the main basis of colinearity loss between cereal genomes.

Discussion

Structural divergence between duplicated genes in plants

Our estimate of the frequency of chromosomal rearrangements (that is, duplicated gene loss) between homoeoalleles in wheat - 54% within less than 1.5 to 3 million years of evolution - needs to be viewed in the context of published studies. Qi *et al.* [22], based on a restriction fragment length polymorphism (RFLP) genotyping approach, mapped 7,104 EST unigenes onto 16,099 loci within the 21 bread wheat chromosomes. Because 39% of the ESTs mapped to the three homoeologous groups, those studies might have suggested that up to 61% of the homoeologs might have lost at least one of the homoeoalleles even despite technological limits due to the RFLP mapping resolution. Overall, we suggest here that, based on our and published data, 54 to 61% (depending on the genetic mapping or chromosome assignation procedures) of the wheat homoeoalleles have been entirely deleted or pseudogenized within less than 3 million years of evolution.

Re-analysis of the paleoduplication within the rice genome, consisting of ten major duplications as part of a WGD event 50 to 70 MYA, has shown that 87.4% of the duplicated genes have lost their orthologous counterparts [10]. Diverged polyploids, such as maize, are likely to have evolved from ancient polyploids by a process of pseudogene formation followed by sequence loss. In a study of the fate of duplicated genes in the maize genome, Lai *et al.* [36] and Messing *et al.* [12] have suggested that, within 5 million years of evolution, about 50% of duplicated genes have been lost through deletion. Nonetheless, gene duplication in maize, *per se*, via (auto)polyploidization may be associated with detectable increases in expression level, as demonstrated by Guo *et al.* [37]. Blanc *et al.* [9] reported similar findings from the *Arabidopsis* genome, where also only 20% of paralogs were retained within duplicated segments. More precisely, the authors stated that 28% and 13.5% of duplicated genes are retained in recent (date back to the *Arabidopsis*/Brassicacae divergence, 24 to 40 MYA) and old (date back to the monocot/dicot divergence, approximately 150 to 200 MYA) duplication blocks, respectively. Considering the recent data obtained in dicots (*Arabidopsis*) and monocots (rice, wheat, maize), our results provide additional support that most of the genetic redundancy originating from polyploidy is erased by a massive loss of duplicated genes by pseudogenization in one of the duplicated segments soon after the polyploidization event.

The structural loss of duplicated genes between paralogous segments as well as gene duplication in tandem (CNVs) accounted for a large part of the erosion of colinearity between cereal genomes. It became clear that using synteny-based approaches to establish a virtual gene order in non-sequenced genomes might mimic up to 77% of the gene order and content [4]. The remaining consists of lineage-specific duplicates loss and CNVs that will not be known until the genome is fully sequenced [4]. However, we can estimate that a large majority of the gene content can be modeled based on synteny, especially to support the development of gene-based markers such as COS [26].

Expression divergence between duplicated genes in plants

As for chromosomal rearrangements, we also need to place our estimate of the frequency of change in expression patterns between homoeoalleles in wheat - 36 to 49% (depending on the considered tissues) within less than 1.5 to 3 million years of evolution - in the context of published studies. Using a similar cDNA-SSCP approach to that reported in this study, Bottley *et al.* [38] demonstrated that for 27% (in leaf) and 26% (in roots) of the considered genes, one homoeologous copy was not detectable within the cDNA samples. Our estimate of functional partitioning between homoeoalleles includes not only a presence/absence variation at the tissue level (49%) but also takes into account the difference in the expression profiles based on a developmental kinetic within a specific organ (36% in wheat grain). Using a cDNA-amplified fragment length polymorphism (AFLP) assay, Kashkush *et al.* [39] estimated that about 5% of the genes are silenced in a newly synthesized allohexaploid, a figure comparable with the study of He *et al.* [40] using a similar approach. This level is substantially lower than our estimates, but not surprising given the time gene silencing could continue over many generations. It certainly confirms that the diploidization process immediately follows the polyploidization event. Exploiting large collections of EST data, Mochida *et al.* [41,42] concluded that silencing affected 11 out of 90 sets of homoeoalleles (12%). Overall, based on our and published data, we suggest that 12 to 49% (depending on the tissues and approaches considered) of the wheat homoeoalleles have been neo- and/or subfunctionalized within less than 3 million years of evolution.

A similar difference between synthetic and ancient hybrids has been demonstrated in cotton, where Adams *et al.* [43,44] used a cDNA-AFLP assay to show that about 5% of all genes are silenced in a newly synthesized allotetraploid, but that about 25% of genes were affected

in natural tetraploid cotton (within 1 to 2 million years of divergence). However, in newly synthesized *A. thaliana* × *A. arenosa* hybrids described by Comai *et al.* [45], only 0.4% of genes are silenced as a direct result of polyploidization, a figure substantially lower than that described in wheat and cotton. On the other hand, a survey on gene expression variation of *A. thaliana* and *A. arenosa*, which split from a common ancestor approximately 1.5 MYA, showed a higher number of approximately 2.5% of gene expression differences. The reason(s) for these disparities remain unclear, but might be a consequence of lower levels of homoeology between the two contributing genomes, and therefore the induction of a lower level of interference in their independent expression. Still, gene silencing certainly appears to be a common phenomenon in established polyploids, and the frequency of silencing seems to increase over time. Using detailed analysis of expression divergence between rice paleoduplicates, Throude *et al.* [10] have shown that 88%, 89%, and 96% diverged in their expression pattern in grain, leaf and root within 50 to 70 million years of evolution, respectively. Blanc *et al.* [9] showed that 57% (for young duplications) to 73% (for old duplications) of paralogs have diverged in expression based on a computational analysis involving 62 Affymetrix microarray experiments in *Arabidopsis*. However, Blanc *et al.* [9] cautioned that the 73% of gene pairs that have diverged in expression in the context of old duplications is an underestimate as cross-hybridization occurred at a high rate in this type of array-based experiment. Finally, expression of maize duplicates has been investigated through EST and cDNA mapping (EST-overgos by Gardiner *et al.* [46] and cDNA-RFLP by Helentjaris *et al.* [47]), suggesting that 20% and 29%, respectively, of the considered probes identified two distinct contigs or loci. These data suggest that 71 to 80% of the maize paralogs have diverged in their expression profiles from both EST and cDNA-based mapping experiments. However, gene silencing of duplicated copies rather than deletion is probably more a gene-dosage effect than just a strict diploidization response. Paralogous copies of prolamin genes (a medium size multigene family) in maize also showed that less than 50% of the duplicated copies remained intact [48]. Interestingly, at the same level, differential gene amplification (such as CNVs) also resulted in subfunctionalization of additional gene copies by divergent transcriptional regulation, mimicking the same events that happen in the same period of evolution between homoeologs [49].

Temporal modeling of structural and expression gene shuffling after duplications in plants

We established clearly in this study that around 70% of homoeoalleles in the hexaploid wheat genomes have been lost (54 to 61%) or have diverged in gene expression (12 to 49%) since 1.5 to 3 MYA. These data confirm and complement the conclusion of Mochida *et al.* [41,42] that, considering 79 genes with scored expression in 10 tissues, 15 (19%) were expressed equally for the three homoeologs whereas the remaining 64 (81%) showed preferential homoeologous gene expression in at least one of the considered tissues. Based on the collective data from wheat detailed in the current article and from other plant species available in the literature, we tried to model the structural and functional consequences of gene set amplification after genome doubling for the last 100 million years of evolution. Figure 5a illustrates plant phylogeny, where speciation events are dated in MYA and known WGDs are marked with red dots. Based on our and other studies from the literature referenced in Table S11 in Additional file 1, we propose that structural rearrangements (from pseudogenization up to deletion) have been suggested to affect 54 to 61% of wheat homoeoalleles (Figure 5b, data point 1), 71 to 80% of maize neoparalogs (Figure 5b, data point 2), 72% of *Arabidopsis* neoparalogs (Figure 5b, data point 3), 87% of rice paleoparalogs (Figure 5b, data point 4), and 86% of *Arabidopsis* paleoparalogs (Figure 5b, data point 5) after 1.5 to 3, 5, 24 to 40, 70 to 100, and 150 to 200 million years of evolution, respectively. Regarding the impact of polyploidy on functional differentiation between duplicated gene copies, our and published data have suggested that 12 to 49% of wheat homoeoalleles (Figure 5b, data point 6), 50% of maize neoparalogs (Figure 5b, data point 7), 57% of *Arabidopsis* neoparalogs (Figure 5b, data point 8), 88% of rice paleoparalogs (Figure 5b, data point 9), and 73% of *Arabidopsis* paleoparalogs (Figure 5b, data point 10) do not exhibit any concerted and redundant expression after 1.5 to 3, 5, 24 to 40, 70 to 100, and 150 to 200 million years of evolution, respectively.

Given the prevalence of gene and genome duplication in the paleohistory of plant, species and lineage development in angiosperms might differ from organisms where genome duplication was rare and where extensive expression divergence following duplication would have a profound impact on the pattern of developmental and regulatory networks. Our data support the idea that after 50 to 70 million years of evolution since grass genomes experienced a shared paleotetraploidization event, the vast majority of the homoeologous genes have been lost within a sister block and that the expression profiles of the remaining gene copies have largely diverged (Figure 5b). Changes in gene expression may have occurred immediately after polyploidy or might need a few generations to reach a new expression status. This trend towards silencing (or gene loss via pseudogenization) or expression shift (via neo- or subfunctionalization) of a

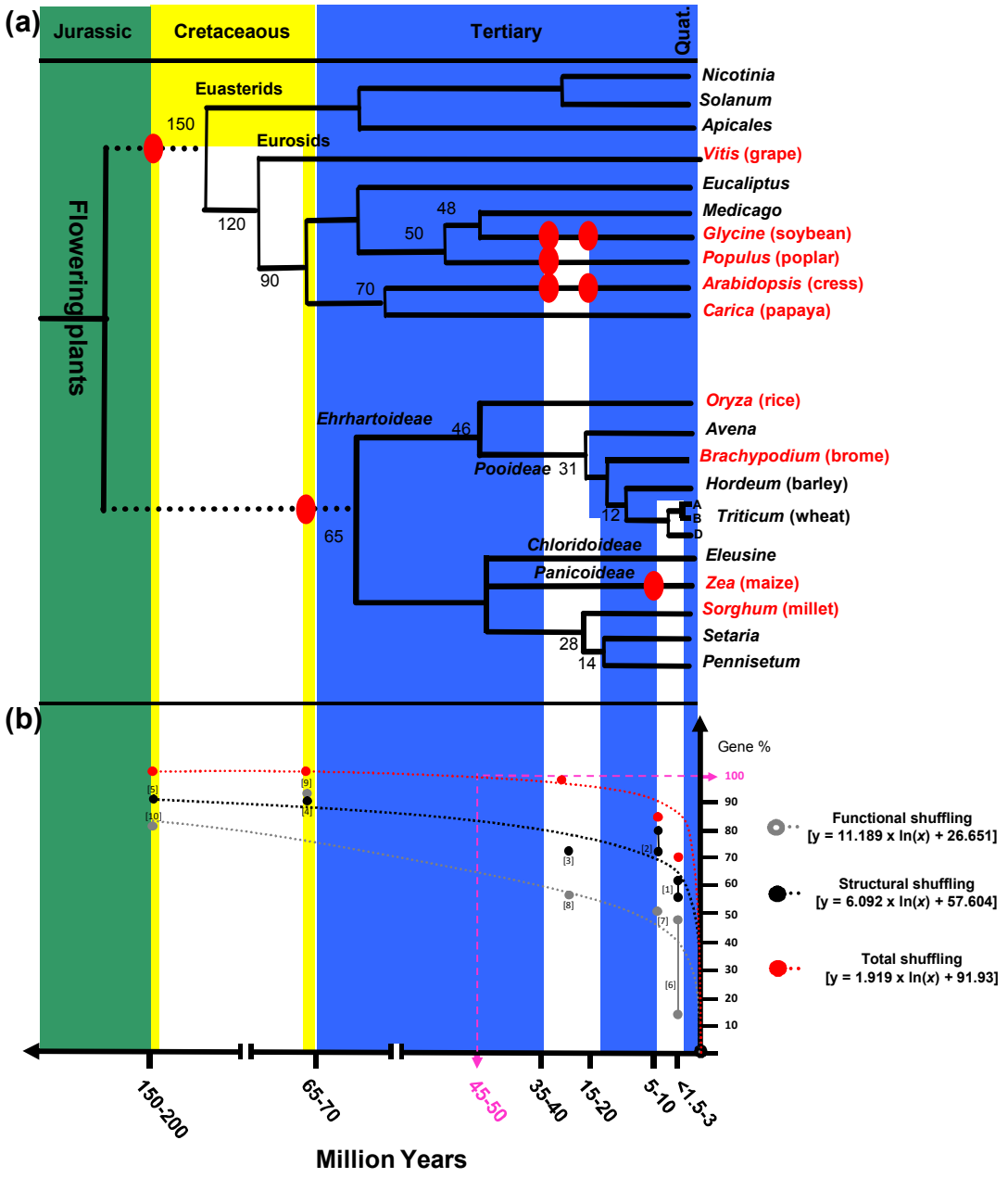


Figure 5 Temporal model of structural and functional gene shuffling in plants. **(a)** Schematic representation of the phylogenetic relationships between angiosperm species. Divergence times from a common ancestor are indicated on the branches of the phylogenetic tree (in million years). Sequenced genomes are indicated in red. WGD events are illustrated with red circles on the tree branches. **(b)** The percentage of total and structural (gene loss) or functional (neo-subfunctionalization) gene shuffling discussed in the article (referred to as data point 'X' in the text) and observed during the last 100 million years of evolution are plotted as dotted red, black and grey logarithmic curves.

particular locus soon after a polyploid event could be advantageous for adaptation and the establishment of a successful polyploid genome compared to its diploid founder progenitor. Figure 5b shows the evolution of gene function (grey dotted curve) and structure (black dotted curve). It follows that loss of duplicated genes due to mutation and deletion appeared to be a rapid and exponential process arising immediately after polyploidy because there is sufficient time for point mutations to accumulate. Moreover, expression modification and silencing of duplicated genes appear to take longer and are probably epigenetically induced (that is, the putative causal factor). Strikingly, based on the deduced total duplicated gene shuffling inference (dotted curves

in Figure 5b) within approximately 10 million years of evolution after a polyploidization event, approximately 50% of the homoeologs have either been lost or sub- or neofunctionalized in plants. The superimposition on these immediate or short-term (putatively mutation-based changes) and longer-term (putatively epigenetic-based changes) responses to genome doubling might explain the observed structural and functional partitioning among gene pairs originating from a duplication event. We can hypothesize that the diploidization that takes place immediately after a WGD is completed for 100% of the duplicated genes after 45 to 50 million years of evolution, the evolutionary timescale necessary to observe that none of the sister duplicated gene copies exhibit any structural or expressional or functional redundancy (pink dotted lines in Figure 5b).

Gene dosage relations, which play a huge role in genome reorganization, are unbalanced after a WGD due to function redundancy between duplicated copies. Structural and functional shuffling occurs relatively soon (within less than a few million years of evolution) after polyploidy in plants that are still cytogenetically polyploidy (such as in the case of bread wheat in the current study), but is still active several million years latter, during or after cytological diploidization (such as in the case of rice, *Arabidopsis* and maize in the current study). We may hypothesize that epigenetic differences between duplicates or even sub-genomes deriving from WGD might have contributed to a gene or genome dominance through the rapid differentiation of expression toward gene dosage balance recovery soon after polyploidy. Wang *et al.* [50] observed silencing of polyploidy-derived duplicates due to hypermethylation in *Arabidopsis* polyploids. Epigenetic mechanisms as well as interaction networks might be the origin of an extremely rapid divergence of expression between duplicated genes soon after polyploidization. It has even been reported that polyploidization-derived modulation of expression between gene pairs was due to epigenetic mechanisms (*sensu lato*) in higher plants (reviewed in [45,51]). Based on our data set and derived conclusions, the bread wheat genome could be considered as a pertinent model for studying the molecular basis of the interaction between homoeologous gene pairs, especially the epigenetic basis of such observed modification in expression between duplicates in response to polyploidizations. The spectrum of phenomena discussed here illustrates the immediate impact of polyploidy on genome structure and its profound implication for evolution. For example, some of the observed genomic changes are known to affect phenotypes in ways that are highly visible to natural selection. A case in point concerns genomics rearrangements that affect the flowering-time locus/network in synthetic *Brassica* polyploids. These polyploidy-induced structural and functional rearrangements may impact traits as relevant as flowering-time divergence in modern plant species.

Conclusions

Even if our estimates of divergence of expression between gene pairs might represent an underestimation of the true values in wheat because the data set is (i) centered on a grain developmental kinetic and then only a sampling of possible environmental conditions or tissues where the duplicated genes may be expressed, and (ii) based on RNA-seq, which may bias low expressed gene and homoeoallele identification, one cannot escape the theme that a large majority of the polyploidy-derived duplicated genes in plants have acquired divergent expression patterns and with them probably functions. Overall, duplication-mediated structural and functional gene shuffling promote a powerful acceleration of evolution in plants.

Materials and methods

Plant material and RNA extraction

Plant material

Two hundred seeds of hexaploid wheat, *Triticum aestivum* (cv. *Récital*), were sown with 4/5 Neuhaus compost and 1/5 Pouzzolane. After 8 weeks of vernalization, plants were transferred to a greenhouse with normalized temperature (approximately 18.5°C), light and hygrometry conditions (60%). The main stem heads were tagged at anthesis and grain samples (endosperm and embryo) were collected at 100, 200, 250, 300 and 500 DD after pollination. Two biological replications of samples were done in 2004 and 2006. Leaves were sampled at different growing stages and pooled, and roots were sampled on 12-day-old seedlings grown in sand.

RNA extraction

Grain wheat (100, 200, 250, 300, 500 DD), root and leaf samples (approximately 1 g of tissue) were ground in liquid nitrogen and extracted with 4.5 ml of buffer (10 mM Tris-HCl, pH7.4, 1 mM EDTA, 0.1 M NaCI, 1% SDS) 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 residues. The nucleic acids were precipitated by addition of 0.1 vol of 3M AcNA pH5.2 and 2 vol of 100% ethanol. After precipitation, RNA was rinsed once with 70% ethanol and the pellets dissolved in RNase-free water. Purification was made with a DNAse treatment RNase-Free DNase Set (Qiagen, [52]) and then an RNeasy MinElute Cleanup Kit (Qiagen). The integrity of RNA was checked with an Agilent 2100 Bioanalyser microfluidics-based platform using a RNA 6000 Nano Chip kit and reagents (Agilent Technologies, [53]).

454 sequencing and cluster assembly

Normalized cDNA library construction

mRNA was purified from 5 μg total RNA by exonuclease digestion followed by LiCl precipitation (mRNA-Only Eucaryotic mRNA Isolation Kit, Epicenter, [54]). mRNA (1 μg) was used for first-strand cDNA synthesis. cDNA synthesis and amplification were done according to the Mint-Universal cDNA Synthesis Kit user manual (Evrogen, [55]). Amplified cDNA (800 ng) was used as starting material in the normalization reaction using the Trimmer Kit (Evrogen), and normalized material was re-amplified for 18 cycles. Normalized cDNA (2 μg) was digested with 10 units SfiI for 2 hours at 48°C. Fragments larger than 800 bp were isolated from a LMP Agarose Gel and purified using the MinElute Gel Extraction Kit (Qiagen). Purified cDNA fragments (200 ng) were ligated to 100 ng using SfiI and dephosphorylated usin pDNR-lib Vector (Clontech, [56]) in 10 μl using a Fast Ligation Kit (NEB, [57]). Ligations were desalted by ethanol precipitation, and re-dissolved in 10 μl water. Three-fold 1.5-μl desalted ligation was used to transform NEB10b competent cells (NEB), and 96 clones were randomly selected for sequencing to verify successful normalization. Roughly a million clones were plated on LB-Cm plates, scraped off the plates and stored as glycerol stocks at -70°C. One half of the cells were used to inoculate a 300-ml Terrific Broth/Cm culture, which was grown for 5 hours at 30°C. Plasmid DNA was prepared using standard methods (Qiagen). Purified plasmid DNA (200 μg) was digested with 100 units SfiI for 2 hours at 48°C. cDNA inserts were gel-purified (LMP-Agarose/MinElute Gel Extraction Kit) and ligated to high-molecular-weight DNA using a proprietary SfiI linker.

Roche 454 FLX library preparation and sequencing of the cDNA concatenates

The five grain samples (100, 200, 250, 300, 500 DD) were equally mixed for sequencing library construction and sequencing with an approximately 30× gene coverage (based on 1 million reads per run and approximately 30,000 expressed genes obtained from the Affymetrix experiment on the same samples discussed in the Results section). Library generation for the 454 FLX sequencing was carried out according to the manufacturer's standard protocols (Roche, [58]). In short, the concatenated inserts were sheared randomly by nebulization to fragments ranging in size from 400 to 900 bp. These fragments were end-polished and the 454 A and B adaptors that are required for the emulsion PCR and sequencing were added to the ends of the fragments by ligation. The resulting fragment library was sequenced on 1 picotiterplate (PTP) on the GS FLX using Roche/454 Titanium chemistry.

Assembly of the sequence reads to transcripts

Prior to assembly the sequence reads were screened for the SfiI linker used for concatenation, the linker sequences were clipped out of the reads and the clipped reads assembled to individual transcripts using the Roche/454 Newbler software (454 Life Sciences Corporation, software release 2.0.01.14) at the following parameter settings: seed step = 12; seed length = 16, minimum overlap length = 40, minimum overlap identity = 90%, alignment identity score = 2, alignment different score = -3. As a consequence, sequence reads were obtained using 454 (Roche) experimental procedures and materials, then sequence clusters were constructed using Newbler Assembler software (release 2.0.01.14) based on a sequence overlap threshold of 40 bases and an identity percentage of a least 90% within overlaps. Sequence clusters were aligned against reference databases for vectors [59], bacterial genomes [60], and mitochondria and chloroplast [61] as well as ribosomal [62] sequences. The 454 sequence data are publicly available at the National Center for Biotechnology Information [63] under accession numbers [JP206682] to [JP238633].

Affymetrix array hybridization and analysis

Hybridization

The Affymetrix [64] wheat GeneChip® oligonucleotide array, which have probes for 55,052 transcripts, was hybridized according to the following procedure. Total RNA (2 μg) from the five grain samples harvested in 2004 (109, 204, 247, 295, 501 DD) and 2006 (125, 186, 231, 292, 489 DD) were used to synthesize biotin-labeled cRNAs with the one-cycle cDNA synthesis kit (Affymetrix). SuperScript II reverse transcriptase and T7-oligo(dT) primers were used to synthesize single-stranded cDNA at 42°C for 1 hour, followed by synthesis of double-stranded cDNA using DNA ligase, DNA polymeraseI, and RNaseH for 2 h at 16°C. After cleaning of the double-stranded cDNA with the Sample Cleanup Module (Affymetrix), *in vitro* transcription was performed in the presence of biotin-labeled UTP using the GeneChip® IVT labeling kit (Affymetrix). The labeled cDNA was purified with the Sample Cleanup Module (Affymetrix) and quantified with RiboGreen RNA quantification reagent (Turner Biosystems, [65]). Fragmentation of 15 μg of labeled cDNA was carried out for 35 minutes at 94°C, followed by hybridization for 16 hours at 45°C to Affymetrix wheat GeneChip® oligonucleotide arrays. After hybridization, the arrays were washed with two different buffers (stringent: 6 SSPE, 0.01% Tween 20; and non-stringent: 100 mm MES, 0.1 M Na^+, 0.01% Tween 20) and stained with a complex solution including Streptavidin R-Phycoerythrin conjugate (Molecular

Probes, [66]) and anti-streptavidin biotinylated antibody (Vector Laboratories, [67]). The washing and staining steps were performed in a GeneChip® Fluidics Station 450 (Affymetrix). The Affymetrix wheat GeneChip® oligonucleotide arrays were finally scanned with the GeneChip® Scanner 3000 7G piloted by GeneChip® Operating software. All these steps were performed on an Affymetrix platform at INRA-URGV in Evry (France).

Statistical data analysis

The raw CEL files were imported in the Bioconductor software package in R for data analysis [68]. The data were normalized with the gcrma algorithm [69] available in the Bioconductor package. To determine differentially expressed genes, we performed a standard two-group *t*-test that assumes equal variance between groups. The variance of the gene expression per group is a homoscedastic variance, where genes displaying extremes of variance (too small or too large) were excluded. The raw *P*-values were adjusted by the Bonferroni method, which controls the familywise error rate [70]. A gene is declared as differentially expressed if the Bonferroni P-value is < 0.05. The raw data are available through the CATdb database (reference AFFY_seed_kinetic_Wheat) [71] and from the Gene Expression Omnibus [72] at the National Center for Biotechnology Information (NCBI), accession number GSE 16457.

cDNA-SSCP primer design and profile analysis

Primer design

Affymetrix wheat GeneChip® sequences were download from the Affymetrix online database [73] and used to design primers pairs. Wheat sequence exons structures were identified through rice/*Brachypodium*/sorghum/maize and wheat sequence alignments and provided to the Primer 3 package to select primer only on one exon using default parameters.

cDNA-SSCP protocol

The absence of contaminating genomic DNA in RNA samples was tested directly by PCR. cDNA was synthesized using Transcriptor First Strand cDNA Synthesis Kit (Roche) and diluted 50 times. PCR products were generated and analyzed with the SSCP protocol according to Quraishi *et al.* [26]. Briefly, cDNA-PCR fragments were produced in two steps. In a total volume of 15 μl, genomic DNA (30 ng) was first amplified with the following PCR mix: 10 mM Tris-HCL, 3.1 mM $MgCl_2$, 50 mM KCl, 0.001% gelatine pH 8.3, 5% glycerol, 400 μM dNTP, 0.4 μM forward and reverse primers, 0.2 U Taq polymerase (Qiagen). This PCR product was diluted (1/10) and re-amplified with the same PCR mix, including 0.2 μM of each labeled primer (6-FAM and NED, Applied Biosystems [74]) in a final volume of 15 μl. The PCR product (2 μl) was then diluted (1/10) and pooled with 0.2 μl of 900 bp MegaBace ET900-R Size Standard (GE Healthcare, [75]), 0.2 μl of 0.3 N NaOH and 9 μl HI-Di Formamide (Applied Biosystems). Fragments were separated by capillary electrophoresis on an ABI3100 (Applied Biosystems) in 50 minutes with a 36 cm capillary. The running polymer consisted of 1× running buffer, 5% Genscan Polymer (Applied Biosystems), 10% glycerol. Samples were denatured for 2 minutes at 95°C and then 10 minutes in ice. The sample buffer consisted of 1× running buffer and 10% glycerol. After denaturation, the samples were injected at 2.5 kV over 50 seconds and separated at 18, 25, and 35°C and 15 kV. Data were analyzed using GeneMapper 3.7 software.

Identification of homeologs, orthologs and paralogs in wheat genomes

The methodology used to reassess the synteny between wheat/rice/*Brachypodium*/sorghum/maize genomes as well as the identification of intra-chromosomal duplications in wheat is described in detail in Salse *et al.* [2,76], Bolot *et al.* [77], and Abrouk *et al.* [4]. Wheat (5,003 mapped unigene set), rice (41,046 genes), *Brachypodium* (25,504 genes), sorghum (34,008 genes) and maize (32,540 genes) genomes were aligned to identify orthologs and co-linear regions [2,3]. Three parameters were used to increase the stringency and significance of BLAST sequence alignment by parsing BLASTn results and rebuilding high scoring pairs (HSPs) or pairwise sequence alignments. The first parameter, aligned length (AL), corresponds to the sum of all HSP lengths. The second, cumulative identity percentage (CIP) corresponds to the cumulative percent of sequence identity obtained for all the HSPs (CIP = Σ nb ID by HSP/AL) × 100). The third parameter is the cumulative alignment length percentage (CALP). It represents the sum of the HSP lengths (AL) for all the HSPs divided by the length of the query sequence (CALP = AL/Query length). The CIP and CALP parameters allow the identification of the best alignment, that is, the highest cumulative percentage of identity in the longest cumulative length, taking into account all HSPs obtained for any pairwise alignment. These parameters were applied to all the BLAST alignments that were performed in the present study. Based on the genome-wide synteny analysis, gene relationships between species are then referenced as COS (for conserved gene pairs), CNV (for tandem duplicated genes), PAV (for non-conserved genes).

Additional material

Additional file 1: Supplemenraty tables. To support the use of the provided RNA-seq data, we provide eleven supplementary tables. Table S1: RNA-sequence quality and coverage features. Table S2: information on the 37,695 wheat sequence clusters. Table S3: the 17,881 homologs between wheat and *Brachypodium*/rice/sorghum/maize genomes. Table S4: the heterologous bread wheat expression map. Table S5: the SSCP

analysis. Table S6: *Brachypodium* heat maps. Table S7: GO classification. Table S8: transcription factor data. Table S9: wheat Affymetrix experiment data. Table S10: the starch pathway analysis. Table S11: the structural/functional shuffling model. The provided supplementary tables provide access to the raw data (gene name, sequence, position, function, expression and statistical data) of the results detailed in the article.

Abbreviations

AFLP: amplified fragment length polymorphism; bp: base pair; CALP: cumulative alignment length percentage; CNV: copy number variation; CDS: coding sequence; CIP: cumulative identity percentage; COS: conserved orthologous set; DD: degree day; EST: expressed sequence tag; GO: Gene Ontology; HSPs: high scoring pairs; MYA: million years ago; RFLP: restriction fragment length polymorphism; RNA-seq: RNA sequencing; SNP: single nucleotide polymorphism; SSCP: Single Strand Conformational Polymorphism; TF: transcription factor; WGD: whole genome duplication.

Acknowledgements

We gratefully acknowledge Joachim Messing (Rutgers, The State University of New Jersey, USA) and Thierry Langin (INRA, Clermont-Ferrand, France) for fruitful discussions in preparing the current article. The authors would also like to thank Isabelle Nadeau, Christine Girousse (INRA Clermont-Ferrand, France) and Biogemma (route d'Ennezat 63720 Chappes, France) for technical support and advice during the plant material preparation. This work has been supported by grants from INRA ('Génétique et Amélioration des Plantes', reference 'Appel d'Offre Transcriptome') and from the 'Agence Nationale de la Recherche' (Program ANRjc-PaleoCereal, reference ANR-09-JCJC-0058-01; program ANR Blanc-PAGE, reference ANR-2011-BSV6-00801).

Author details

[1]INRA, UMR 1095, Genetics, Diversity and Ecophysiology of Cereals, 234 avenue du Brézet, 63100 Clermont-Ferrand, France. [2]INRA, Unité de Recherches en Génomique Végétale, 2 rue Gaston Crémieux, CP 5708, F-91057 Evry Cedex, France.

Authors' contributions

CP designed the experiment, performed the analysis and participated in manuscript preparation. FM performed the bioinfomatic analysis and participated in manuscript preparation. Carole Confolent performed the molecular biology experiments. SB performed the transcriptome (Affymetrix) analysis. JS designed the research program, managed the research group and wrote the article.

Received: 20 May 2011 Revised: 25 August 2011
Accepted: 2 December 2011 Published: 2 December 2011

References

1. Ohno S: *Evolution by Gene Duplication* Berlin: Springer-Verlag; 1970, 160.
2. Salse J, Abrouk M, Murat F, Quraishi UM, Feuillet C: **Improved standards and new comparative genomics tools provide new insights into grasses paleogenomics.** *Brief Bioinf* 2009, **10**:619-630.
3. Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C: **Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals.** *Proc Natl Acad Sci USA* 2009, **106**:14908-14913.
4. Abrouk M, Murat F, Pont C, Messing J, Jackson S, Faraut T, Tannier E, Plomion C, Cooke R, Feuillet C, Salse J: **Palaeogenomics of plants: synteny-based modelling of extinct ancestors.** *Trends Plant Sci* 2010, **15**:479-487.
5. Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF: **Evolutionary genetics of genome merger and doubling in plants.** *Annu Rev Genet* 2008, **42**:443-461.
6. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics.** *Proc Natl Acad Sci USA* 2004, **101**:9903-9908.
7. Davis JC, Petrov DA: **Do disparate mechanisms of duplication add similar genes to the genome.** *Trends Genet* 2005, **21**:548-551.
8. Ganko EW, Meyers BC, Vision TJ: **Divergence in expression between duplicated genes in Arabidopsis.** *Mol Biol Evol* 2007, **24**:2298-2309.
9. Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.** *Plant Cell* 2004, **16**:1679-1691.
10. Throude M, Bolot S, Bosio M, Pont C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, Murigneux A, Charmet G, Perez P, Salse J: **Structure and expression analysis of rice paleo duplications.** *Nucleic Acids Res* 2009, **37**:1248-1259.
11. Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J: **On the tetraploid origin of the maize genome.** *Comp Funct Genomics* 2004, **5**:281-284.
12. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA: **Sequence composition and genome organization of maize.** *Proc Natl Acad Sci USA* 2004, **101**:14349-14354.
13. Haberer G, Hindemitt T, Meyers BC, Mayer KF: **Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of Arabidopsis.** *Plant Physiol* 2004, **136**:3009-3022.
14. Fawcett JA, Maere S, Van de Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event.** *Proc Natl Acad Sci USA* 2009, **106**:5737-5742.
15. Van de Peer Y, Maere S, Meyer A: **The evolutionary significance of ancient genome duplications.** *Nat Rev Genet* 2009, **10**:725-732.
16. Donoghue PC, Purnell MA: **Genome duplication, extinction and vertebrate evolution.** *Trends Ecol Evol* 2005, **20**:312-319.
17. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C: **Major ecological transitions in wild sunflowers acilitated by hybridization.** *Science* 2003, **301**:1211-1216.
18. Hegarty M, Hiscock S: **Polyploidy: doubling up for evolutionary success.** *Curr Biol* 2007, **17**:927-929.
19. Bicknell RA, Koltunow AM: **Understanding apomixis: recent advances and remaining conundrums.** *Plant Cell* 2004, **16**:228-245.
20. Feldman M, Lupton FGH, Miller TE: **Wheats.** In *Evolution of Crops..* 2 edition. Edited by: Smartt J, Simmonds NW. London: Longman Scientific; 1995:184-192.
21. Nesbitt M, Samuel D: **From staple crop to extinction? The archaeology and history of the hulled wheats.** In *Proceedings of the First International Workshop on Hulled Wheats: 21-22 July 1995; Castelvecchio Pascoli, Tuscany, Italy. Volume 4.* Edited by: Padulosi S, Hammer K, Heller J. Biodiverstiy International; 1996:41-100, Promoting the Conservation and Use of Underutilized and Neglected Crops.
22. Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorák J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin , Gustafson JP, Conley EJ, *et al*: **A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat.** *Genetics* 2004, **168**:701-712.
23. Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J: **Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.** *Genome Res* 2010, **20**:1545-1557.
24. Mayer KF, Martis M, Hedley PE, Simková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubaláková M, Suchánková P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Dolezel J, Waugh R, Stein N: **Unlocking the barley genome by chromosomal and comparative genomics.** *Plant Cell* 2011, **23**:1249-1263.
25. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.** *Plant Cell* 2010, **22**:1686-1701.
26. Quraishi UM, Abrouk M, Bolot S, Pont C, Throude M, Guilhot N, Confolent C, Bortolini F, Praud S, Murigneux A, Charmet G, Salse J: **Genomics in cereals: From genome-wide conserved orthologous set (cos) sequences to candidate genes for trait dissection.** *Funct Integr Genomics* 2009, **9**:473-484.
27. International Brachypodium Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon*.** *Nature* 2010, **463**:763-768.

28. Xu JH, Messing J: **Diverged copies of the seed regulatory Opaque-2 gene by a segmental duplication in the progenitor genome of rice, sorghum, and maize.** *Mol Plant* 2008, **1**:760-769.
29. Wan Y, Poole RL, Huttly AK, Toscano-Underwood C, Feeney K, Welham S, Gooding MJ, Mills C, Edwards KJ, Shewry PR, Mitchell RA: **Transcriptome analysis of grain development in hexaploid wheat.** *BMC Genomics* 2008, **9**:121.
30. Sulpice R, Pyl ET, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques MC, Von Korff M, Steinhauser MC, Keurentjes JJ, Guenther M, Hoehne M, Selbig J, Fernie AR, Altmann T, Stitt M: **Starch as a major integrator in the regulation of plant growth.** *Proc Natl Acad Sci USA* 2009, **106**:10348-10353.
31. Zhu T, Budworth P, Chen W, Provart N, Chang HS, Guimil S, Su W, Estes B, Zou G, Wang X: **Transcriptional control of nutrient partitioning during rice grain filling.** *Plant Biotechnol J* 2003, **1**:59-70.
32. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps.** *Genome Res* 2008, **18**:1944-1954.
33. Seoighe C, Gehring C: **Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome.** *Trends Genet* 2004, **20**:461-464.
34. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution.** *Genome Biol* 2006, **7**:R43.
35. Wu X, Qi X: **Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication.** *BMC Evol Biol* 2010, **10**:145.
36. Lai J, Ma J, Swigonová Z, Ramakrishna W, Linton E, Llaca V, Tanyolac B, Park YJ, Jeong OY, Bennetzen JL, Messing J: **Gene loss and movement in the maize genome.** *Genome Res* 2004, **14**:1924-1931.
37. Guo M, Davis D, Birchler JA: **Dosage effects on gene expression in a maize ploidy series.** *Genetics* 1996, **142**:1349-1355.
38. Bottley A, Xia GM, Koebner RMD: **Homoeologous gene silencing in hexaploid wheat.** *Plant J* 2006, **47**:897-906.
39. Kashkush K, Feldman M, Levy AA: **Gene loss, silencing and activation in a newly synthesized wheat allotetraploid.** *Genetics* 2002, **160**:1651-1659.
40. He P, Friebe BR, Gill BS, Zhou JM: **Allopolyploidy alters gene expression in the highly stable hexaploid wheat.** *Plant Mol Biol* 2003, **52**:401-414.
41. Mochida K, Yamazaki Y, Ogihara Y: **Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags.** *Mol Genet Genomics* 2003, **270**:371-377.
42. Mochida K, Kawaura K, Shimosaka E, Kawakami N, Shin-I T, Kohara Y, Yamazaki Y, Ogihara Y: **Tissue expression map of a large number of expressed sequence tags and its application to *in silico* screening of stress response genes in common wheat.** *Mol Genet Genomics* 2006, **276**:304-312.
43. Adams KL, Cronn R, Percifield R, Wendel JF: **Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing.** *Proc Natl Acad Sci USA* 2003, **100**:4649-4654.
44. Adams KL, Percifield R, Wendel JF: **Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid.** *Genetics* 2004, **168**:2217-2226.
45. Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B: **Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids.** *Plant Cell* 2000, **12**:1551-1568.
46. Gardiner J, Schroeder S, Polacco ML, Sanchez-Villeda H, Fang Z, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, Chou H, Wing R, Soderlund C, Coe EH Jr: **Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization.** *Plant Physiol* 2004, **134**:1317-1326.
47. Helentjaris T, Weber D, Wright S: **Identification of the genomic location of duplicate nucleotide sequences in maize by the analysis of restriction fragment length polymorphisms.** *Genetics* 1998, **118**:353-363.
48. Song R, Messing J: **Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize.** *Plant Physiol* 2002, **130**:1626-1635.
49. Song R, Llaca V, Linton E, Messing J: **Sequence, regulation, and evolution of the maize 22-kD alpha zein gene family.** *Genome Res* 2001, **11**:1817-1825.
50. Wang X, Shi X, Hao B, Ge S, Luo J: **Duplication and DNA segmental loss in the rice genome: implications for diploidization.** *New Phytol* 2005, **165**:937-946.
51. Chen ZJ, Ni Z: **Mechanisms of genomic rearrangements and gene expression changes in plant polyploids.** *Bioessays* 2006, **28**:240-252.
52. **Qiagen Company..** [http://www.qiagen.com].
53. **Agilent Company..** [http://www.home.agilent.com].
54. **Epicenter Company..** [http://www.epibio.com].
55. **Evrogen Company..** [http://www.evrogen.com].
56. **Clontech Company..** [http://www.lablife.org].
57. **NEB Company..** [http://www.neb.com].
58. **Roche Company..** [http://www.roche.com].
59. **VECTOR DB..** [ftp://ftp.ncbi.nih.gov/pub/UniVec/UniVec].
60. **Sanger Institute: Escherichia coli..** [http://www.sanger.ac.uk/resources/downloads/bacteria/escherichia-coli.html].
61. **ORGANELLE DB..** [http://organelledb.lsi.umich.edu].
62. **Silva..** [http://www.arb-silva.de].
63. **National Center for Biotechnology Information..** [http://www.ncbi.nlm.nih.gov].
64. **Affymetrix Company..** [http://www.affymetrix.com].
65. **Turner Biosystems Company..** [http://www.topac.com].
66. **Molecular Probes Company..** [http://www.invitrogen.com].
67. **Vector Laboratories Company..** [http://www.vectorlabs.com].
68. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
69. Irizarry RA, Ooi SL, Wu Z, Boeke JD: **Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants.** *Stat Appl Genet Mol Biol* 2003, **2, Article1.**
70. Dudoit S, Gentleman RC, Quackenbush J: **Open source software for the analysis of microarray data.** *Biotechniques* 2003, , **Suppl:** 45-51.
71. Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V: **CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform.** *Nucleic Acids Res* 2008, **36**:D986-990.
72. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets - 10 years on.** *Nucleic Acids Res* 2011, **39**:1005-1010.
73. **Affymetrix: Wheat..** [http://www.affymetrix.com/Auth/analysis/downloads/data/wheat.consensus.zip].
74. **Applied Biosystems Company..** [http://www.appliedbiosystems.com].
75. **GE healthcare Company..** [http://www.gehealthcare.com].
76. Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C: **Identification and characterization of conserved duplications between rice and wheat provide new insight into grass genome evolution.** *Plant Cell* 2008, **20**:11-24.
77. Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J: **The 'inner circle' of the cereal genomes.** *Curr Opin Plant Biol* 2009, **12**:119-125.

doi:10.1186/gb-2011-12-12-r119
Cite this article as: Pont *et al.*: RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.). *Genome Biology* 2011 **12**:R119.

3. Discussion

3.1. Utilisation des espèces apparentées pour étudier la régulation du génome du blé tendre hexaploïde ; la recherche translationnelle.

Dans le cadre des travaux précédents publiés en 2011, les séquences parcellaires des génomes des blés diploïdes, tétraploïdes et hexaploïdes n'étaient pas disponibles (*cf.* Figure 6, page 9). Nous avons mis en œuvre une approche de génomique translationnelle consistant à utiliser les génomes de *Brachypodium*, du riz, du sorgho et du maïs comme référence, pour l'étude du transcriptome du grain chez le blé tendre. A l'initiation de ces travaux et dans l'état des connaissances en 2011, le répertoire des transcrits homéologues chez le blé n'était que très peu décrit et ceci uniquement par des approches 'microarray' (Wan *et al.* 2008). Les pseudomolécules 'blé' n'étant pas disponible, nous avons utilisé les génomes modèles de *Brachypodium*, riz, sorgho et maïs pour ordonner les transcrits, modéliser les sept groupes chromosomiques du blé et *in fine,* étudier l'expression des copies homéologues. Ces résultats *in silico* ont été validés par une approche 'RNA-SSCP' assignant les copies homéologues par PCR sur les 21 chromosomes du blé tendre (pour un lot de 100 transcrits). Nous avons choisi d'étudier Recital, une variété 'hiver' à fort rendement et à forte valeur boulangère. Nous nous sommes focalisés sur le développement du grain, tissu de choix dans l'étude du rendement et valorisable pour la qualité boulangère. Cet article a permis (1) de caractériser le transcriptome du grain de blé en développement par l'identification et la localisation *in silico* de 37 695 transcrits, et (2) de quantifier l'asymétrie expressionnelle des homéologues chez le blé tendre. En réponse à la néo-hexaploïdie, la divergence d'expression des copies homéologues héritées, est de 36 à 49 % selon le tissu considéré. La divergence des copies dupliquées héritée de la paléo-tétraploïdie est, elle, de l'ordre de 80 %.

Dans cette étude, j'ai réalisé une analyse approfondie du transcriptome, *via* cinq échantillons et deux réplicats biologiques, couvrant le développement du grain à l'aide de puces à ADN GeneChip® (technologie affymetrix). Cette technologie présente de nombreux avantages, notamment sa densité (plus de 1,3 millions de cibles, représentant 55 052 transcrits de blé), ainsi que sa spécificité d'hybridation due à des sondes très courtes (25 pb). L'analyse a permis d'identifier plus de 27 000 gènes exprimés au cours des 5 stades, et 2 901 gènes différentiellement exprimés. Toutefois, cette approche n'est pas exhaustive car le set de gènes présent sur la puce est loin d'être complet et peut amener à une sous-estimation du catalogue génique.

J'ai ainsi approfondi cette étude *via* l'approche RNA-seq. Les cinq échantillons (100, 200, 250, 300 et 500 degrés jours après floraison (dd)) couvrant le développement du grain ont été mélangés de façon équimolaire, puis séquencés. J'ai choisi une approche de séquençage dite 'nouvelle génération' (NGS) générant de grands fragments (plateforme GS 454, roche) afin de pouvoir séparer les différents transcrits homéologues avec de faibles variations de séquence (homéoSNPs). Ainsi, nous avons caractérisé le répertoire des transcrits du développement du grain. Dans notre étude, ce répertoire est constitué de 37 695 séquences correspondant à 38 % du contenu génique identifié à ce jour (99 386 gènes modèles selon Bolser *et al.* 2014). Depuis, ce répertoire de transcrits spécifique 'développement du grain' a été complété en 2014, à travers le séquençage de 30 échantillons, selon la méthode illumina paired-end, aboutissant à 46 487 gènes exprimés (Pfeifer *et al.*2014). Cette légère

différence en termes de catalogue en gènes exprimés, est due au nombre conséquent d'échantillons dans l'étude de Pfeifer *et al. (*2014) ainsi qu'à une dissection fine des différents tissus qui permet de séparer la couche aleurone de l'endosperme et gagner ainsi en exhaustivité.
Dans notre étude Pont *et al.* (2011), nous montrons qu'à travers 7 158 gènes (conservés chez les céréales apparentées), seulement 193 sont associés à l'expression des 3 homéologues chez le blé et donc, traduisant une redondance des transcrits. Cet effectif traduit une très faible redondance expressionnelle pour moins de 3 % des triplets A, B et D chez le blé tendre. Même si l'assemblage des gènes, réalisé ici à l'aide des génomes d'espèces apparentées, a pu fusionner des séquences d'homéologues à forte similarité (et ainsi surévaluer cet effectif), une forte divergence d'expression apparaît pour une très grande majorité des triplets. A noter que la sous- et néo-fonctionnalisation ne peuvent être distinguées à ce stade. Afin de valider ce résultat, j'ai montré, par assignation PCR d'un lot de 100 gènes, que 54 % des gènes ont perdu structuralement une copie, et 49 % montrent au moins une copie 'silencée' dans l'un des 3 tissus (36% dans le grain en développement). La méthode RNA-SSCP (Single Strand Conformation Polymorphism) utilisée dans l'étude, permet la séparation des copies homéologues selon leurs polymorphismes de séquence. La détection n'étant pas quantitative, seules les copies totalement absentes ont été comptabilisées et décrites comme réprimées (ou 'silencées' pour reprendre le terme utilisé dans la littérature). Par cette approche, l'analyse des copies perdues et silencées dans le grain, démontre que seulement 34 % des triplets montrent une redondance parfaite d'expression entre les copies A, B et D. Le reste, 66 %, répond au phénomène de néo- ou sous-fonctionnalisation déjà décrit, additionné à la perte structurale (délétion d'un des homéologues). Au regard de la WGD datant de 90 MYA, seulement 19,3 % des paralogues sont encore présents et exprimés (par comparaison avec le chromosome 2 de *Brachypodium* et les chromosomes 1 et 3 de blé). Nos travaux ainsi que ceux de la littérature montrent très clairement que les modifications d'expressions entre copies dupliquées sont massives après un événement de polyploïdie et sont induites relativement rapidement après un tel choc génomique.

3.2. L'apport des nouvelles séquences génomiques.

Un nouveau transcriptome - A l'époque de la rédaction de cet article, en 2011, nous n'avions pas recherché de corrélation entre la divergence d'expression et un effet de dominance des sous-génomes (décrit dans le Chapitre I), ou une quelconque compartimentation génomique. Cette analyse mérite désormais d'être réalisée grâce au synténome de blé, délivrant un répertoire de 72 900 gènes localisés sur les 21 chromosomes (*cf. Chapitre 1.2.3)*. Pour cela, j'ai entrepris le séquençage (RNA-seq) des 3 stades de développement du grain décrits dans l'article précédent (division cellulaire à 100dd, remplissage du grain à 250dd et dessiccation à 500dd), avec la technologie illumina pour plus de profondeur de lectures (*cf.* Figure 41). Obtenant plus de 60 millions de lectures par échantillon, Moaine El Baidouri (post-doctorant de l'équipe) a réalisé le mapping (*via* Bowtie2; Langmead and Salzberg 2012), l'analyse d'expression (*via* le calcul du RPKM : nombre de lecture du gène/ (nombre total de lectures de l'échantillon * taille du transcrit)), ainsi que le clustering en modules d'expression du développement du grain (*via* WGCNA, Langfelder *et al.* 2008 ; *cf.* Figure). L'analyse des données RNA-seq sur la base des 99 386 gènes blé de référence (IWGSC 2014, Bolser *et al.* 2014), montre que 28 776 gènes sont exprimés. J'ai utilisé ces données par la suite, en exploitant 20 067 gènes exprimés (RPKM>1) au cours du

développement du grain et positionnés sur le synténome (72 900 gènes ordonnés sur les 21 chromosomes parmi les 99 386 gènes disponibles). Cet effectif correspond à 27,5 % du répertoire en gène du synténome (et à 72 % des transcrits inventoriés dans l'article Pont *et al.* 2011). J'ai ainsi pu actualiser nos connaissances de l'asymétrie expressionnelle des copies homéologues chez le blé tendre à différentes échelles : génome, sous-génomes et gènes.

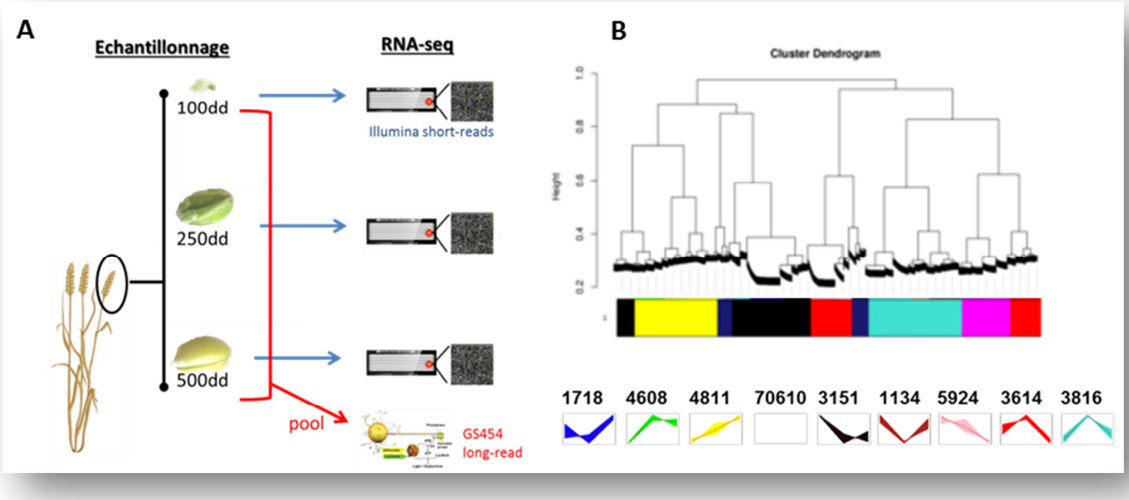


Figure 41. Illustration de la méthodologie d'analyse des données RNA-seq du grain en développement chez le blé
(A) Schématisation de l'approche développée dans la nouvelle étude RNA-seq. Dans l'étude précédente (Pont *et al.* 2011) trois stades de développements du grain ont été mélangés pour obtenir des séquences relativement longues avec la technologie GS454 (matérialisée en rouge). Dans cette nouvelle études, les 3 stades ont été repris séparément (matérialisés en bleu) pour être séquencés en profondeur, avec la technologie illumina. Les lectures mappées peuvent être comptabilisées et normalisées par le calcul du RPKM (nombre de lecture du gène/ (nombre total de lectures de l'échantillon * taille du transcrit)). dd : degree day. (B) Analyse des transcrits RNA-seq avec le logiciel WGCNA (Langfelder *et al.* 2008). Les gènes sont groupés selon leur corrélation d'expression et un dendrogramme est obtenu par clustering hiérarchique. Les couleurs montrent l'assignation aux 9 modulent générés qui sont représentés en dessous avec les 3 stades (le module 'white' correspond aux gènes non exprimés). Les effectifs en gènes sont reportés au-dessus de chaque module. *Source : M. El Baidouri.*

Analyse à l'échelle des gènes exprimés **-** J'ai utilisé le lot de 8 671 triplets (homéologues A, B et D) précédemment décrits (*cf.* chapitre 1.2.3) dont 5 573 sont positionnés sur les 21 chromosomes (*via* le synténome) et 3 741 sont exprimées au cours du développement du grain. Dans un premier temps, j'ai comparé les copies simplement exprimées (RPKM>1) versus les copies non transcrites (RPKM<1). Cette comparaison révèle le taux de néo ou sous-fonctionnalisation des gènes par silencing ou activation. La proportion des gènes sous- et néo-fonctionnalisés est de l'ordre de 55% des triplets exprimés dans le grain, et semble ne pas être aléatoire (*cf.* Figure 42). J'ai pu noter que si l'on considère les triplets avec seulement 1 copie exprimée sur les 3 (soient 881 triplets), le génome A est le plus exprimé en nombre de gènes (*cf.* Figure ; #343) et le génome B, le plus exprimé en intensité d'expression (*cf.* Figure 42 ; RPKMmoy=13.9). Si l'on considère les triplets avec seulement 1 copie réprimée ou 'silencée', le génome B est plus fréquemment touché par la diploïdisation expressionnelle avec 39% des cas, la copie réprimée est retrouvée plus fréquemment sur le sous-génome B (*cf.* Figure ; #354 pour B ; #285 pour A et #269

pour D) dit sensible. A noter que l'on retrouve seulement 1 478 triplets exprimés au sein des 3 sous-génomes à la fois, soit 17% des gènes au regard du grain en développement.

L'hybridation homoploïde du génome D implique l'apparentement structural des copies tripliquées de la manière suivante : B(AD)>A(BD)>>D(AB). Pour rappel ; D(AB)=19 %, B(AD)=42 % et A(BD)=38 % (*cf.* Chapitre I, page 43). Est-ce que l'apparentement structural des copies homéologues se traduit par une similarité de profil d'expression ? Il ne semble pas être le cas ; l'apparentement expressionnel AB n'est pas aussi faible qu'attendu, avec une expression concertée D(AB)=30 %, B(AD)=39 %, et A(BD)=31 % (*cf.* Figure). Cette observation démontre que l'expression des copies semble être indépendante de l'apparentement structural des copies homéologues. Si l'on considère les triplets pour lesquels un des homéologues est 'silencé', il apparaît que la copie ayant perdu son expression provient plus fréquemment du sous génome B ; les copies les plus proches sont A et D (B(AD)=39% des cas). L'hybridation homoploïde serait trop ancienne pour trouver sa trace au niveau expressionnel.

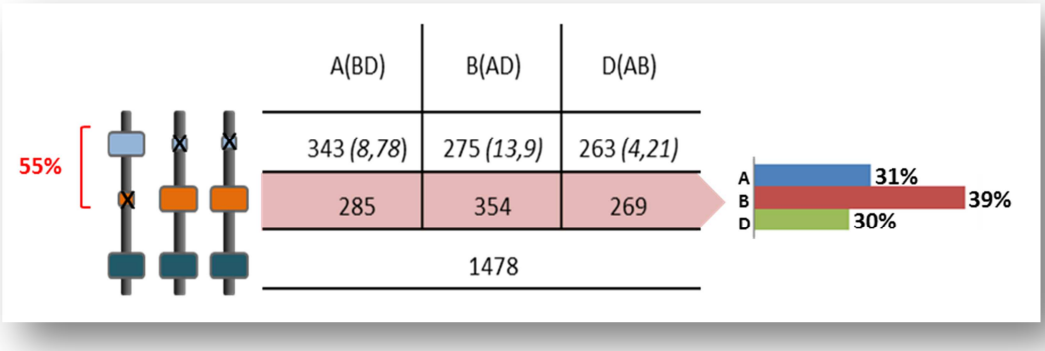


Figure 42. Répartition des copies néo- sous-fonctionnalisées parmi les triplets de gènes exprimés dans le grain
A gauche, est représenté le fractionnement génique ; 1 à 3 copies exprimées, mais avec les 3 copies présentes structuralement dans chaque cas. A droite pour chaque cas, l'effectif des gènes est reporté, spécifiquement exprimés (gros rectangle) dans un des trois génomes (ligne 1, rectangles bleus), dans deux génomes communément (ligne 2, rectangles oranges), puis dans les trois (ligne 3, rectangles verts). Lorsqu'une seule copie est exprimée (ligne 1), le RPKM moyen des 3 stades est mentionné entre parenthèses, démontrant l'intensité plus importante portée par les singletons du génome B. Le pourcentage des gènes sous- et néo-fonctionnalisés est de l'ordre de 55% des triplets exprimés.

Comme nous l'avons vu dans le chapitre I (*cf.* Figure 22, page 26), chez le blé hexaploïde la compartimentation de l'expression des gènes peut être appréhendée en comparant les gènes conservés entre les céréales au sein de 6 régions (3 sous-génomes homéologues x 2 paléo-sous-génomes). A partir des 20 067 gènes exprimés (RPKM>1) au cours du développement du grain de blé, j'ai étudié les différences d'expression entre ces 6 compartiments. L'analyse globale, grâce au synténome, permet d'ordonner les gènes, visualiser leur répartition et ainsi, observer qu'elle n'est pas homogène au sein des chromosomes (*cf.* Figure 43B). En effet, les gènes présents sur les compartiments sensibles (*cf.* Figure 43A ; histogrammes rouges) sont légèrement plus exprimés que sur les compartiments dominants (*cf.* Figure 43A ; histogrammes bleus), avec des valeurs de RPKM 1,31 fois plus élevées (p=0.006, T-test unilatéral par paires). Ce phénomène a été retrouvé chez le riz, *Brachypodium* et le sorgho (données de M. El Baidouri). Chez le maïs, à l'opposé, une surexpression a été observée sur les sous-génomes dominants (Schnable *et al.* 2011). Chez le blé, si cette analyse est réalisée avec l'ensemble

des gènes (99 386), et sans considérer uniquement les gènes conservés chez les céréales (*i.e.* le syntenome), aucune asymétrie d'expression entre les 6 sous-génomes n'est détectée comme dans l'étude de Pfeifer *et al.* 2014. Chez le blé, la néo-dominance seule n'est que très peu visible en terme de divergence d'expression (entre les compartiments A, B et D ; *cf.* Figure 43A ; histogrammes à droite). Il n'y a pas d'effet graduel de divergence ou de plasticité expressionnelle comme tel était le cas au niveau structural (*cf.* Figure 33B, page 42). Il peut être suggéré que le mécanisme impliqué dans cette diploïdisation expressionnelle agit au niveau des gènes (différence observée entre homéologues) mais peu au niveau de blocs chromosomiques (faible différence observée entre sous-génomes).

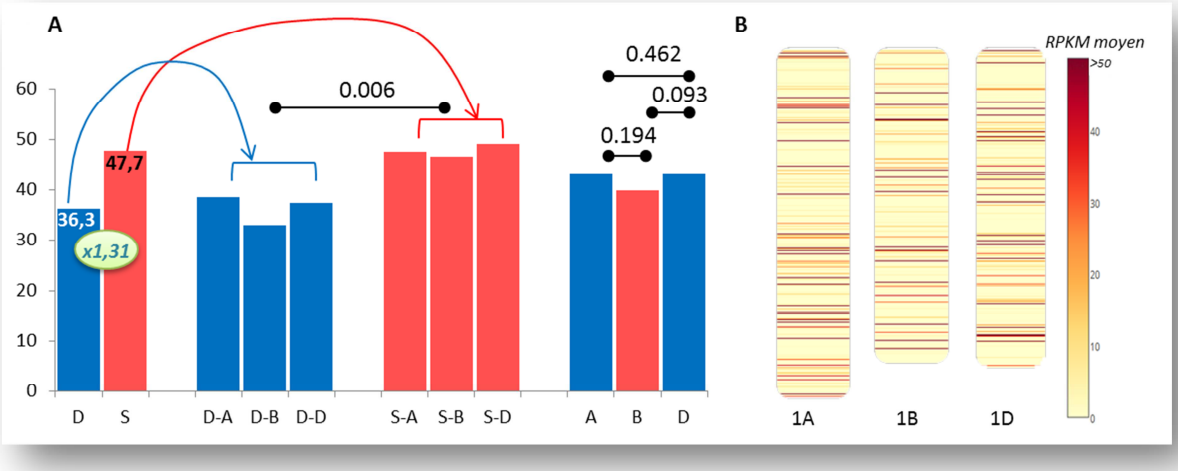
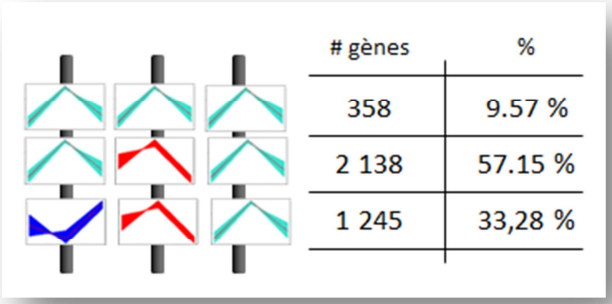


Figure 43. Etude la compartimentation de l'expression des gènes chez le blé hexaploïde
(A) Niveau d'expression (RPKM moyen des 3 stades) des triplets ABD selon leur compartimentation ancestrale (paléo-sous-génome D et S), récente (sous génomes A, B et D) et en 6 compartiments (3 sous-génomes homéologues x 2 paléo-sous-génomes ; S-A, S-B, S-D, D-A, D-B et D-D). Les p-values sont mentionnées au-dessus des histogrammes (T-test unilatéral par paires). (B) Représentation des gènes exprimés selon leur niveau d'expression (RPKM moyen de 3 stades), ordonnés sur les chromosomes et selon le synténome. Exemple des chromosomes 1A, 1B et 1D.

Au niveau de l'analyse GO de ces gènes, il existe un enrichissement significatif ($p<.001$) entre les différents compartiments D et S. Le compartiment dominant est associé à des gènes exprimés de type 'signal transduction', 'small GTpase meditated signal transduction', 'signaling', 'cellular homeostasis' ; des fonctions biologiques nécessaires pour maintenir un équilibre ou activer des processus cellulaires à partir de signaux transmis. Le compartiment sensible, quant à lui, est associé à des gènes exprimés relevant des GO impliqués dans la régulation des processus biologiques 'biological regulation', 'regulation of cellular process'. On retrouve ici également beaucoup de processus nécessaires à la biosynthèse des protéines, ce qui est l'objectif premier du grain en développement : 'translation', 'translation elongation', 'protein localization', 'maintenance of protein location', 'biosynthetic process'… La plasticité apportée par le compartiment sensible peut ainsi permettre la néo- et sous-fonctionnalisation des gènes impliqués dans les processus de régulation. Le niveau d'expression plus élevé de ce compartiment peut apporter un gain aux éléments régulés en aval.

Analyse à l'échelle des profils d'expression du développement du grain **-** J'ai pu affiner mon analyse en me focalisant sur la concertation expressionnelle des triplets en considérant 3 stades de développement du grain. J'ai là encore, utilisé le lot de 8 671 triplets précédemment décrits (*cf.* chapitre 1.2.3) dont 5 573 sont positionnés sur les 21 chromosomes (*via* le synténome) et 3 741 sont exprimés. Ces derniers ont donc pu être groupés selon leur corrélation d'expression dans 9 modules d'expressions (*cf.* Figure 41, page 59). Ainsi, lorsque je considère les 3 stades de développement du grain, seulement 9,57 % des triplets homéologues ont un même profil d'expression (appartenance aux mêmes modules, *cf.* Figure 44). Toujours sur ce même jeu de données, 2 138 triplets ont 2 copies concertées avec le même profil d'expression (*cf.* Figure 44, ligne 2), soit 57,15 % des triplets exprimés (38 % du répertoire). 1 245 triplets possèdent 3 copies divergentes, soit 33,28 % de néo- et sous-fonctionnalisation (22 % du répertoire).



# gènes	%
358	9.57 %
2 138	57.15 %
1 245	33,28 %

Figure 44. Néo et sous-fonctionnalisation des triplets homéologues dans le grain en développement.
A gauche, est représenté le profil d'expression de copies homéologues (à titre d'exemple 3 modules sont matérialisés). A droite, l'effectif et la proportion pour chaque cas sont notés. Les modules d'expression représentent le profil RPKM à travers les 3 stades de développement du grain. La première ligne correspond à 3 copies redondantes, la 2ème, à une copie néo ou sous-fonctionnalisée et la 3ème ligne correspond à une diploïdisation fonctionnelle totale (3 copies dans 3 modules d'expression distincts).

Au niveau de l'analyse GO, il existe un enrichissement significatif entre les triplets montrant une redondance d'expression et ceux néo-sous-fonctionnalisés. L'enrichissement est relativement diversifié (FDRs>0,05) ; 42 processus biologiques sont enrichis ainsi que 32 fonctions moléculaires pour les copies montrant une redondance d'expression. Pour les copies néo- et sous-fonctionnalisées, 24 processus biologiques sont enrichis ainsi que 52 fonctions moléculaires (*cf.* Figure 45). Pour les copies à expression concertée, les GO associées concernent des fonctions cellulaires de base, la catalyse de réactions, l'interaction avec le cytosquelette ou encore la conservation de l'intégrité du ribosome. Pour les copies néo- et sous-fonctionnalisées, les gènes enrichis ont pour but de moduler l'activité d'enzymes et de faciliter la diffusion de molécules à travers les membranes (*cf.* Figure 45).

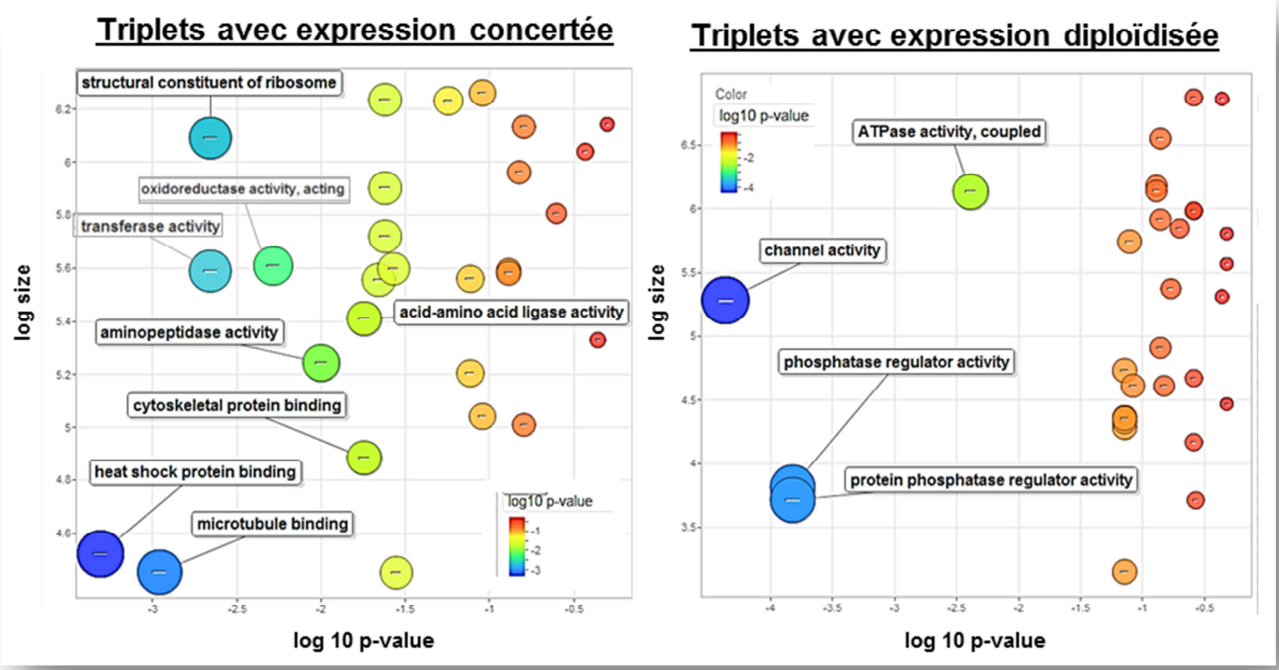
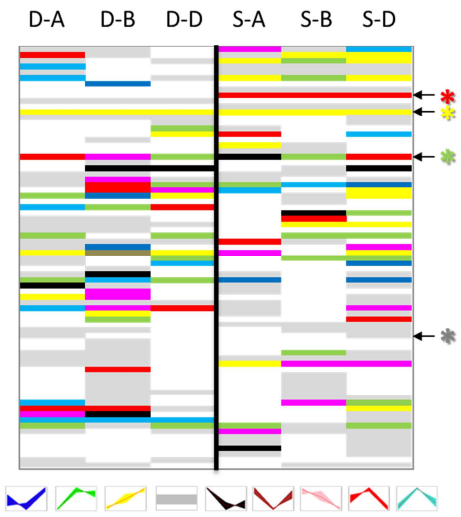


Figure 45 : Description ontologique des triplets associés à une expression concertée et diploïdisée.
Enrichissement en fonctions moléculaires des triplets avec trois copies exprimées de façon concertées (à gauche) et non concertées (à droite). Dans ce dernier cas, une seule des 3 copies présentes structuralement est exprimée, soit une diploïdisation expressionnelle pour les 2 autres copies. L'analyse statistique et la visualisation sont réalisées avec les outils agriGO, et revigo en filtrant les gènes avec un FDR<0,05. L'axe des abscisses représente le log 10 de la p-value et l'axe des ordonnées indique la fréquence du terme GO dans la base de données blé. L'enrichissement est calculé ici par rapport aux 3 741 triplets exprimés et localisés sur les 21 chromosomes. La couleur et la taille des cercles est proportionnelle à la p-value, selon l'échelle du graphique.

Analyse des gènes ohnologues - Enfin, je me suis focalisée sur l'analyse des copies ancestrales datant de plus de 90 MYA conservés donc, en 6 copies, chez le blé tendre (on parle d'ohnologues). Ces 6 copies sont présentes dans les 6 compartiments (3 sous-génomes homéologues x 2 paléo-sous-génomes ; S-A, S-B, S-D, D-A, D-B et D-D) et peuvent être résistantes à la diploïdisation après les cycles récurrents de WGD. J'ai pu identifier seulement 75 groupes de gènes conservés en 6 copies, dont 45 ont au moins une copie exprimée (60%). Ces ohnologues sont donc en effectif limité dû à la très forte diploïdisation du génome moderne (*cf.* Figure 46A, étoile grise et rouge). Là encore, une différence de niveau d'expression entre ohnologues D et ohnologues S est observable avec un facteur x1,47 en faveur du bloc sensible (RPKM moyen de 36,70 pour le compartiment S et 25,01 pour D ; p-value=0,00219). Au regard de ces 45 ohnologues exprimés, seulement 1 ohnologues (*cf.* Figure 46A, étoile jaune) conserve le même profil d'expression de ses 6 copies au cours du développement du grain (même module). D'autre part, 9 ohnologues (20 %) ont conservé un même profil d'expression (ancestral) pour au moins une copie homéologues (A ou B ou D) dans chaque sous-génome (D ou S, *cf.* Figure 46A, étoile verte). Les copies dupliquées après 90 MYA conservent donc l'expression ancestrale à hauteur de 20% ; au moins 2 copies sur 6 ont la même expression, avec une sur la région D et une sur la région S.

L'analyse GO (*cf*. Figure 46B) montre que les ohnologues exprimés sont enrichis en 10 processus biologiques (P), 6 fonctions moléculaires (F) et 0 en composant cellulaire. Les processus retrouvés impliquent la biosynthèse de macromolécules, l'interaction sélective et de manière non-covalente (binding), l'incorporation de fer et de soufre exogène dans une molécule et des cofacteurs nécessaires pour l'activité des enzymes. Ici, ces gènes ont été conservés en plusieurs copies potentiellement fonctionnelles sur plus de 90 MYA car elles doivent jouer un rôle essentiel pour la survie de l'espèce. Dans un tissu tel que le grain, ces fonctions biologiques peuvent être associées à l'accumulation des réserves du grain, avec notamment de nombreuses protéines soufrées. Il est cohérent que de telles fonctions soient alors conservées ancestralement (les ohnologues relèvent d'une duplication y a 90 MYA) car le remplissage de la graine assure sa fertilité et le maintien de l'espèce. En effet, on observe que 9% des gènes de la catégorie 'iron-sulfur cluster assembly' sont présents sous forme d'ohnologues exprimés dans le grain. Ce processus apparaît ainsi vital ou 'primaire' pour la plante au point de le conserver en 6 copies redondantes après 90 MYA d'évolution.

A



B

GO term	Ontology	Description	Number in input list	Number in Ref	p-value	FDR
GO:0016226	P	iron-sulfur cluster assembly	6	68	1,90E-15	6,00E-14
GO:0051188	P	cofactor biosynthetic process	6	914	7,90E-09	1,60E-07
GO:0044249	P	cellular biosynthetic process	15	18890	2,20E-08	3,40E-07
GO:0051186	P	cofactor metabolic process	6	1258	5,10E-08	6,40E-07
GO:0009058	P	biosynthetic process	15	20626	7,20E-08	7,60E-07
GO:0006412	P	translation	6	4563	8,10E-05	7,20E-04
GO:0010467	P	gene expression	9	14438	3,10E-04	2,10E-03
GO:0009059	P	macromolecule biosynthetic process	9	14608	3,40E-04	2,10E-03
GO:0034645	P	cellular macromolecule biosynthetic process	9	14577	3,40E-04	2,10E-03
GO:0044237	P	cellular metabolic process	15	49113	3,90E 03	2,20E 02
GO:0051540	F	metal cluster binding	6	564	4,60E-10	6,70E-09
GO:0051536	F	iron-sulfur cluster binding	6	564	4,60E-10	6,70E-09
GO:0005506	F	iron ion binding	6	3588	2,10E-05	2,10E-04
GO:0005488	F	binding	21	78438	5,80E-04	4,20E-03
GO:0003676	F	nucleic acid binding	9	18400	1,90E-03	1,10E-02
GO:0046914	F	transition metal ion binding	6	10265	5,60E-03	2,70E-02

Figure 46. Expression et analyse GO des ohnologues au sein des 6 compartiments.
(A) Représentation de l'expression des copies ohnologues répartis en 9 modules d'expression. Les ohnologues sont potentiellement présents en 6 copies dans les 6 compartiments (3 sous-génomes homéologues x 2 paléo-sous-génomes ; S-A, S-B, S-D, D-A, D-B et D-D). Les couleurs montrent l'assignation de gènes aux 9 modules d'expression qui sont représentés dessous avec 3 stades (le module 'gris' correspond aux gènes non exprimés). Les cases blanches correspondent aux copies non présentes (perdues). (B) Enrichissement de la GO (Gene Ontology). L'analyse montre que les ohnologues exprimés sont enrichis en 10 processus biologiques (P) et 6 fonctions moléculaires (F). L'effectif enrichi dans chaque classe, ainsi que l'effectif de la classe sont mentionnés dans la table. Les valeurs de p-value et FDR sont mentionnées dans les colonnes à droite.

Les différentes analyses réalisées sur l'asymétrie d'expression des sous-génomes chez le blé se sont basées sur l'analyse de l'expression des gènes chez l'espèce hexaploïde naturelle sans prendre en compte les progéniteurs diploïdes et tétraploïdes. Les différences observées sont-elles dues à la

polyploïdisation ? Ou sont-elles le reflet de divergences déjà présentes chez ces progéniteurs ? Les travaux qui ont débuté dans l'équipe à la suite des résultats que j'ai obtenus visent à répondre à cette question.

4. Perspectives ; l'étude de blés synthétiques

Les travaux précédents ont démontré une forte asymétrie d'expression des gènes homéologues. Toutefois ces analyses ont été réalisées *via* l'étude de l'expression des gènes chez le blé hexaploïde seulement. Ces divergences d'expression sont-elles héritées des parents progéniteurs ou dues à la polyploïdisation ? Pour aller plus loin, il semblerait intéressant de prendre en considération l'expression des géniteurs dans la comparaison des profils d'expressions du blé, à l'image de ce qui a été étudié chez le coton (Chaudhary *et al.* 2009). En effet, la néo-fonctionnalisation s'illustre par un gain d'expression dans la descendance (par exemple) et la sous-fonctionnalisation par une compartimentation des deux profils parentaux (ou ancestraux) selon les conditions spacio-temporelles (*cf.* Figure 47).

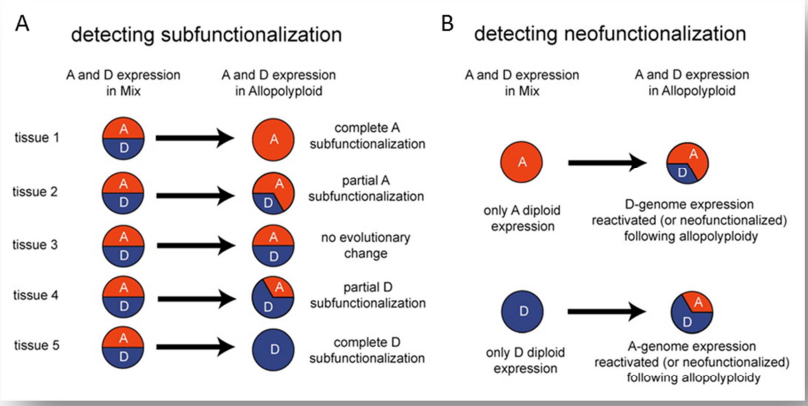


Figure 47. Exemple de détection de néo- et sous-fonctionnalisation chez le coton allotétraploïde.
(A) Détection de la sous-fonctionnalisation de gènes à partir du profil d'expression des géniteurs diploïde A et D ; un des deux profils ancestraux est retenu majoritairement selon les tissus. (B) Détection de la néo-fonctionnalisation de gènes ; un cumul des deux profils ancestraux est retrouvé dans le coton allotétraploïde à la différence de son géniteur. *Source : Chaudhary et al. 2009.*

L'asymétrie d'expression peut être caractérisée en se focalisant sur la provenance des copies exprimées ou ELD 'expression level dominance' (Yoo *et al.* 2013). Ainsi, chez le coton (AADD), l'étude spécifique d'hybrides interspécifiques F1, des allopolyploïdes synthétiques et des naturels comparés aux géniteurs a permis de mesurer l'étendue et la direction de l'ELD. Une dominance 'ELD' est observée en faveur du sous-génome A, sauf chez l'allopolyploïde synthétique où la dominance est en faveur du sous-génome D (Yoo *et al.* 2013). Ainsi, il semble que la mise en place de la dominance soit immédiate suite à l'hybridation et paraît s'accentuer au cours du temps puisque la répression d'une des copies est plus forte chez les allopolyploïdes naturels que chez les synthétiques et hybrides F1.

A la lecture des récents travaux publiés chez le coton, un design d'expérience a été mis en place chez le blé synthétique afin de mieux caractériser la divergence d'expression des copies redondantes des sous-

génomes. Pour cela, des lignées synthétiques (néo-polyploïdes) issues du croisement réciproque entre Langdon (AABB) x tauschii-109 (DD) peuvent être étudiées pour comparer l'expression des gènes entre les polyploïdes AABBDD (naturels et synthétiques) et les deux progéniteurs AABB et DD. Ces plantes ont été cultivées en chambre contrôlée en 2015 et 2016 et soumises à un stress thermique (de l'ordre de +8°C au début du développement du grain, *cf.* Figure 48). Les échantillons prélevés couvrent la cinétique du développement du grain déjà étudiée dans les travaux présentés dans ce chapitre (division cellulaire à 100dd, remplissage du grain à 250dd et dessiccation à 500dd). Ce matériel permettra d'analyser l'impact du génome paternel et maternel dans l'établissement de la dominance des sous-génomes.

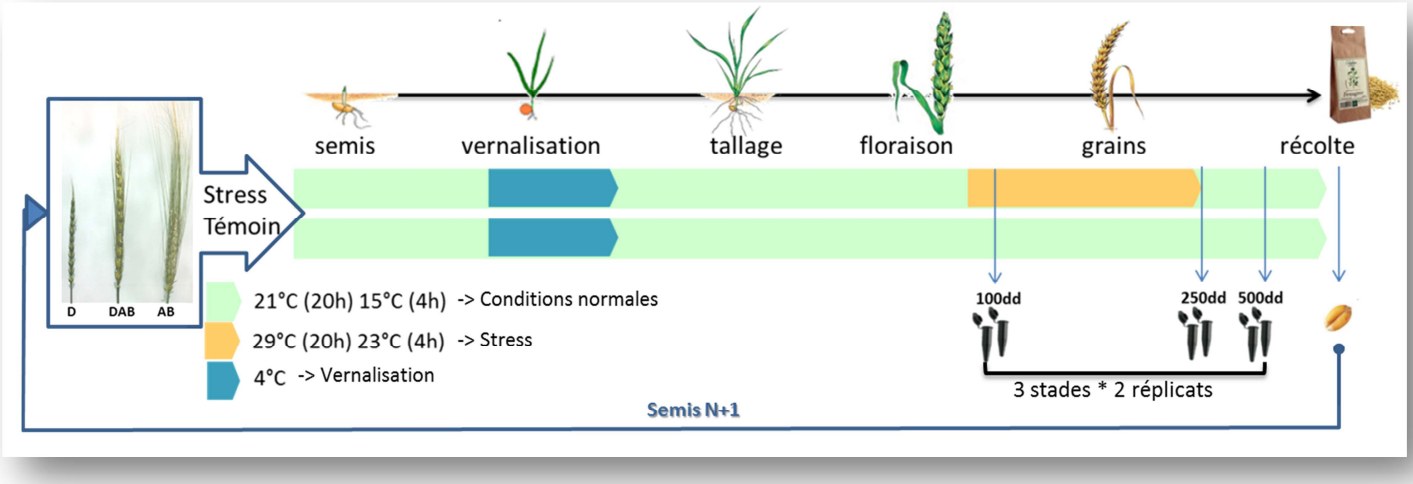


Figure 48. Design expérimental de l'étude de la dominance des sous-génomes sur un modèle de blé néo-synthétique
Les plantes diploïdes, tétraploïdes et hexaploïdes (illustrées à gauche) sont cultivées en chambres contrôlées et soumises à un stress de +8°C (en orange). Au minimum, 2 prélèvements (répliquas) sont réalisés par stade.

Ce dispositif permet de quantifier les modifications 'omiques' en réponse à la polyploïdisation et également en réponse au stress. Nous parlons au sens large de modifications 'omiques' car nous prévoyons le séquençage de ces échantillons en RNA-seq (expression des gènes), smallRNA (expression des petits ARN), BS-seq (méthylation de l'ADN), ATAC-seq (régions transcriptionnellement actives), Chip-seq (avec la marques d'histone H3K9ac associées à la transcription des gènes) et HiC (conformation 3D de l'ADN), et ceci pour 60 échantillons (2 conditions, 3 stades avec 2 réplicats biologiques, 2 blés synthétiques, une variété cultivée et les 2 parents (soit 3 niveaux de ploïdie (D, AB et ABD)). Selon Doyle *et al.* 2008, des processus épigénétiques (sans modification de séquence) *via* la méthylation de l'ADN, la modification des histones et le silencing de gènes par de petits ARN interférents, peuvent également être impliqués dans le phénomène de diploïdisation observé. Il semble que l'allopolyploïdie induit des changements de méthylation chez les spartines au niveau des transposons (Parisod *et al.* 2009b) et plus important, au niveau des gènes, chez le blé (Chagué *et al.* 2010). Les variations 'omiques' (RNA-seq, smallRNA, BS-seq, ATAC-seq et HiC) entre les homéologues A, B et D chez l'hexaploïde devront être

comparées à la somme des gènes parentaux qui modélise l'additivité (AABBxDD=AABB+DD) et pourront être alors classifiés selon les catégories définies en Figure 49 selon Yoo *et al.* 2013 et li *et al.* 2014. En effet, si pour un gène donné, la valeur de son allèle au sein du blé hexaploïde ABD montre une aussi forte valeur que son parent diploïde D (à l'opposé de l'autre parent), l'ELD associée sera en faveur d'une dominance en provenance du parent D. La polyploïdie peut également introduire un phénomène de transactivation ou répression d'une copie (silencing) en raison d'un nouvel environnement de régulation du noyau polyploïde (*cf.* Figure 49 ; li *et al.* 2014).

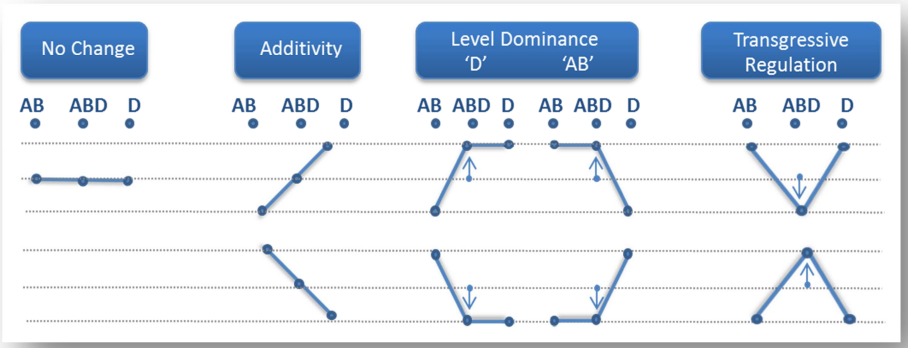


Figure 49. Modélisation des différents profils d'expression d'un génome blé hexaploïde, relatif à un parent tétraploïde AB et diploïde D.
Le niveau d'expression est schématisé et peut suivre ou non la valeur de ses parents ; la dominance observable est matérialisée par les flèches (*Adaptée de Yoo et al. 2013.). Trois classes d'expression sont définies : No Change, Additivity, Dominance, Transgressive.*

Quelques études ont été menées sur des blés nouvellement synthétisés (Chagué *et al.* 2010, Guo *et al.* 2014, li *et al.* 2014). Dans ces études, une analyse du transcriptome a été menée soit par puce ou par séquençage. Elles montrent que l'allopolyploïdie s'accompagne d'une altération modérée de la méthylation des éléments transposables et impacterait les méthylations héritables au niveau des gènes. Chagué *et al.* 2010 ont observé un biais vers l'un des génomes parentaux concernant majoritairement les homéoallèles du génome D, mais qui n'est pas le même chez tous les blés synthétiques étudiés. Il semblerait que les changements de méthylation observés chez les spartines affectent le plus souvent le génome maternel (Parisod *et al.* 2009).

Suite à l'analyse détaillée de l'expression des copies redondantes (prenant compte de l'analyse ELD), il serait intéressant de corréler ces résultats avec la fonction des gènes. Ainsi, les gènes montrant un effet de dominance parentale pourront être associés à une fonction biologique précise. Pour pousser plus en avant l'analyse des gènes les plus impactés par une dominance parentale (ou dominance des sous-génomes), les gènes seront classés en 3 catégories : triplicats ABD, duplicats (AB ; AD et DB), singletons, mais aussi les gènes spécifiques de l'espèce blé (c.à.d non orthologues). Il sera alors possible d'observer si l'enrichissement pour certains types de gènes existe en fonction de leur nombre de copies dans le génome.

5. Conclusion

Mes travaux décrits dans ce chapitre ont permis de démontrer une différenciation massive de l'expression des gènes dupliqués chez le blé tendre moderne. Au regard de la polyploïdie et au sein des 3 génomes ; seulement 9,57 % des gènes ont un même profil d'expression et 33,28 % ont 3 copies divergentes après moins de 500 000 ans, date de la première hybridation du blé lors de la tétraploïdie. De même, si on s'intéresse à la duplication ancestrale il y a 90 MYA, on parle alors de gènes ohnologues, 20 % au moins ont conservé un profil ancestral.

Au niveau fonctionnel, j'ai montré que la polyploïdie chez le blé génère la spécialisation expressionnelle des gènes dupliqués dans de très fortes proportions. Le blé hexaploïde montre une très forte spécialisation des copies (diploïdisation fonctionnelle) sur les chromosomes dits sensibles. Quel est l'impact de la diploïdisation structurale (Chapitre I) ainsi que la diploïdisation expressionnelle (Chapitre II) des copies redondantes de gènes retenus sur l'élaboration des phénotypes chez le blé tendre ?

CHAPITRE 3

Impact de la polyploïdisation sur les caractères phénotypiques

1. Introduction à l'étude de l'impact de la polyploïdie sur la capacité adaptative du blé tendre

Nous avons établi dans les précédents chapitres que la polyploïdie chez le blé était source de plasticité génomique au niveau structural et fonctionnel. Quel est, aujourd'hui, l'impact de cette plasticité héritée d'événements datant de plus de 10 000 ans sur les phénotypes ou caractères agronomiques travaillés en sélection? Dans la nature, la diversité au sein des populations permet aux espèces d'évoluer dans une grande variété de milieux et favorise leur adaptation dans des environnements contrastés. Le blé tendre est cultivé dans une très grande gamme de milieux. En effet, il a été domestiqué il y a 10 000 ans en Irak (Croissant fertile) et depuis, il est cultivé à la fois dans de hautes latitudes (France, Canada, Ukraine…), et à la fois dans les pays du sud avec des variétés plus résistantes à la sécheresse (au Maroc par exemple).

La diploïdisation structurale et fonctionnelle est-elle un moteur pour l'élaboration de tels caractères phénotypiques ? Existe-t-il un lien entre plasticité, dominance génomique et caractères phénotypiques ? Dans les deux premiers volets de cette thèse, j'ai montré que le mécanisme de pertes de gènes ou d'accumulation de mutations est non aléatoire chez le blé tendre. J'ai identifié des blocs chromosomiques (sous-génomes) dominants avec une forte stabilité, et des blocs plus sensibles avec une forte plasticité structurale. Au niveau fonctionnel, j'ai montré également que la polyploïdie chez le blé génère la spécialisation expressionnelle des gènes dupliqués dans de très fortes proportions, avec une plus forte expression des gènes ancestraux portés par les sous-génomes sensibles. L'hypothèse qui découle de mes précédents travaux, serait que l'accumulation de mutations plus fréquentes sur le génome sensible après la polyploïdisation, associée à une plus forte expression des gènes, participerait à l'apparition de nouveaux phénotypes. Certains compartiments génomiques seraient donc potentiellement propices à l'élaboration de nouveaux phénotypes *et in fine* favoriseraient l'adaptation de la plante à son environnement contraint.

Dans un premier temps, et sur la base de la dominance structurale et expressionnelle des sous-génomes chez le blé tendre, nous souhaitons étudier la diploïdisation phénotypique. A l'instar de la perte de la redondance des gènes et de leur expression entre les sous-génomes du blé tendre, peut-on mettre en évidence une diploïdisation phénotypique ? Les gènes/loci responsables de caractères agronomiques ne seraient alors pas portés par les trois sous-génomes A, B et D. On parlera ici d'asymétrie phénotypique ou de diploïdisation des caractères.

1.1. Asymétrie génomique des caractères agronomiques majeurs

De nombreux exemples sont retrouvés dans la littérature avec des gènes contrôlant des caractères phénotypiques ou agronomiques majeurs chez le blé tendre et qui seront présentés dans ce chapitre.
La dureté du grain (gènes Pin) - A titre d'exemple, le locus HA, comme précédemment décrit (*cf.* chapitre I, 1.3 Asymétrie génique des sous-génomes, Figure 25, page 30), illustre le lien entre organisation du génome et phénotype. Le caractère majeur de dureté du grain, permettant de produire du pain à partir de la farine de blé tendre, est diploïdisé. En effet, les gènes *pina* et *pinb* sont portés

uniquement par le chromosome 5D, après la perte des copies homéologues (5A et 5B) lors de l'événement de tétraploïdie.

La taille de la plante (gènes Rht) - Un autre exemple de diploïdisation de caractère majeur est illustré par le locus Rht, contrôlant la hauteur de la plante. Ce locus a été largement sélectionné après la guerre et a joué un rôle majeur lors de la révolution verte (résistance à la verse, facilitation de la récolte, et augmentation du nombre de grains par épis). Selon Pearce *et al.* 2011, les allèles présents sur les sous-génomes B et D (gènes Rht-B1b et Rht-D1b) génèrent une réduction de 20% de la taille de la plante, dues à des substitutions introduisant un codon stop.

La floraison (gène VRN1) - Le gène vrn1 (vernalisation) est également un exemple de diploïdisation de caractère adaptatif majeur : la floraison. En effet, différentes mutations et délétions au sein du promoteur, ou des introns des gènes vrn (Fu *et al.* 2005) conduisent à l'apparition d'un allèle dominant, VRN1 'printemps' en comparaison de l'allèle sauvage récessif, vrn1 'hiver'. Les individus portants ces mutations ont la capacité de fleurir sans un passage au froid levant l'état végétatif (Loukoianov *et al.* 2005). Ces mutations ont donc le pouvoir de raccourcir le cycle de développement de la plante (qui met plus rapidement en place ses organes reproducteurs) et de s'adapter au climat chaud et sec. A l'inverse, les variétés hivers résistent au froid. Cette variabilité de précocité a permis d'élargir les zones de production du blé tendre. Toutefois, l'effet des trois homéologues (vrn1A, vrn1B, vrn1D) n'est pas équivalent dans la réponse à la vernalisation avec un fort effet de l'allèle VRN1A (Fu *et al.* 2005).

La teneur en protéines du grain (gènes Glu) - Si l'on considère les gluténines de haut poids moléculaire, représentant 10 % des protéines de réserve du grain, chaque sous-génome ne joue pas un rôle équivalent dans l'accumulation de ces protéines de réserve. Sur 109 lignées de blé hexaploïde étudiées par Feldman *et al.* 2012, presque 40 % possèdent deux sous-unité Glu-A1, 40 % possèdent une seule sous unité Glu-A1 et 20 % des lignées n'ont pas le locus Glu-A1. Parallèlement, il a été démontré que dans 60 % des lignées étudiées, l'allèle Glu-A1-2 est inactif bien qu'il soit actif chez le blé diploïde parental (Feldman *et al.* 2012). De plus, un polymorphisme plus important a été observé sur le locus 1B par rapport à celui du 1A, une tendance retrouvée également sur les gliadines de haut poids moléculaire (Galili et Feldman, 1983).

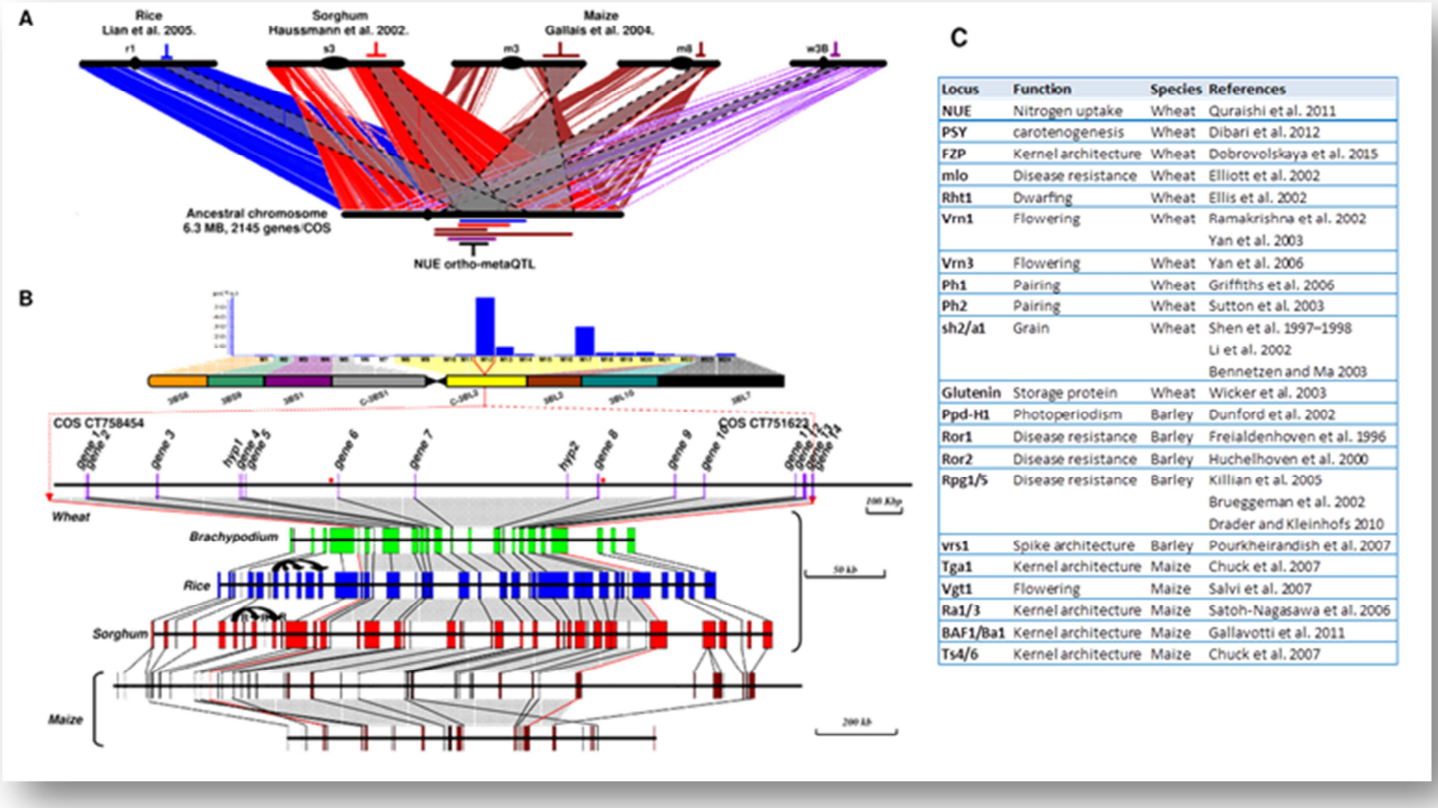
Les précédents exemples illustrent le continuum entre plasticité génomique et phénotypique ou l'asymétrie entre les sous-génomes du blé tendre a conduit à une diploïdisation phénotypique de telle sorte que les sous-génomes ne jouent pas le même rôle dans l'établissement de caractères agronomiques majeurs tels que la dureté du grain, la taille de la plante, la floraison, la teneur en protéines du grain.

1.1. Diploïdisation des caractères agronomiques quantitatifs

Chez le blé tendre, la majorité des caractères agronomiques quantitatifs (QTLs) retrouvés dans la littérature ne sont portés que par un locus très spécialisé, illustrant la diploïdisation des caractères quantitatifs au-delà de caractères agronomiques majeurs présentés dans la section précédente.

L'assimilation de l'azote - Les travaux de l'équipe auxquels j'ai participé sur l'analyse des déterminants géniques de l'assimilation de l'azote, ont permis d'identifier un QTL sur le chromosome 3B, expliquant 8.8% de la variation du caractère au sein de 3 populations chez le blé tendre. Grâce à la génomique

translationnelle, les informations connues sur la fonction biologique d'un gène provenant d'une espèce modèle apparentée, comme le riz, permettent d'interpréter la fonction d'un gène conservé chez le blé. La Figure 50A-B illustre l'utilisation de cette approche de génomique translationnelle chez le blé pour ce caractère contrôlé par l'enzyme glutamate synthase (GoGAT) conservée chez les céréales (Quraishi *et al.* 2011). Les QTL identifiés (*cf*. Figure 50A) chez le maïs, le riz et le sorgho, contrôlés par cette même enzyme GoGAT, ont pu être projetés sur le synténome de blé à l'aide de marqueurs COS cartographiés. Ces QTL co-localisent avec le QTL du chromosome 3B du blé tendre (*cf*. Figure 50B), faisant du gène GoGAT le candidat de ce QTL. Le gène GoGAT est présent sur les deux autres sous-génomes (3A et 3D) mais une différence d'expression expliquerait la diploïdisation de ce caractère, où uniquement l'allèle présent sur le chromosome 3B serait impliqué dans l'assimilation de l'azote. A l'instar du gène GOGAT, cette approche de génomique translationnelle, à partir des céréales apparentées, a été mise en œuvre pour de multiples caractères portés par un seul sous-génome chez le blé tendre hexaploïde (*cf*. Figure 50C).



C

Locus	Function	Species	References
NUE	Nitrogen uptake	Wheat	Quraishi et al. 2011
PSY	carotenogenesis	Wheat	Dibari et al. 2012
FZP	Kernel architecture	Wheat	Dobrovolskaya et al. 2015
mlo	Disease resistance	Wheat	Elliott et al. 2002
Rht1	Dwarfing	Wheat	Ellis et al. 2002
Vrn1	Flowering	Wheat	Ramakrishna et al. 2002 Yan et al. 2003
Vrn3	Flowering	Wheat	Yan et al. 2006
Ph1	Pairing	Wheat	Griffiths et al. 2006
Ph2	Pairing	Wheat	Sutton et al. 2003
sh2/a1	Grain	Wheat	Shen et al. 1997–1998 Li et al. 2002 Bennetzen and Ma 2003
Glutenin	Storage protein	Wheat	Wicker et al. 2003
Ppd-H1	Photoperiodism	Barley	Dunford et al. 2002
Ror1	Disease resistance	Barley	Freialdenhoven et al. 1996
Ror2	Disease resistance	Barley	Huchelhoven et al. 2000
Rpg1/5	Disease resistance	Barley	Killian et al. 2005 Brueggeman et al. 2002 Drader and Kleinhofs 2010
vrs1	Spike architecture	Barley	Pourkheirandish et al. 2007
Tga1	Kernel architecture	Maize	Chuck et al. 2007
Vgt1	Flowering	Maize	Salvi et al. 2007
Ra1/3	Kernel architecture	Maize	Satoh-Nagasawa et al. 2006
BAF1/Ba1	Kernel architecture	Maize	Gallavotti et al. 2011
Ts4/6	Kernel architecture	Maize	Chuck et al. 2007

Figure 50. Génomique translationnelle chez le blé tendre pour la dissection du QTL d'assimilation de l'azote sur le chromosome 3B

(A-B) Clonage du gène GOGAT contrôlant le locus NUE (Nitrogen Use Efficiency) chez le blé. (A) Projection des QTLs conservés *via* la synténie entre les chromosomes de blé (3B à droite), de maïs (m3 et m8), de sorgho (s3) et de riz (r1). Les QTLs sont représentés sur chaque espèce et projetés sur l'ancêtre (bas). (B) L'intervalle du QTL est matérialisé par l'histogramme bleu (LOD score) sur le chromosome 3B. La microsynténie (blé, *Brachypodium*, riz, sorgho, maïs) de la région est visualisée à travers 14 gènes (rectangles) portés par un même BAC du chromosome 3B du blé tendre. Le gène GoGAT est le gène 14, en rouge, conservé entre les espèces et colocalisant avec le QTL NUE. *Source Quraishi et al. 2011.*

Le rendement et la qualité - Une méta-analyse de 297 QTLs chez le blé (travaux de thèse Quraishi *et al.* 2009 ; Quraishi *et al.* 2011) a permis d'étudier la redondance phénotypique au sein des trois génomes. A partir de 10 populations, 22 Meta-QTLs (*i.e.* QTLs pour un même caractère provenant d'au moins deux populations distinctes) ont été identifiés pour le rendement, la teneur en protéines du grain et la valeur boulangère. La localisation fine des régions impliquées montre une complète diploïdisation de ces caractères ; les QTL sont présents uniquement sur l'un des sous-génomes (*cf.* Figure 51).

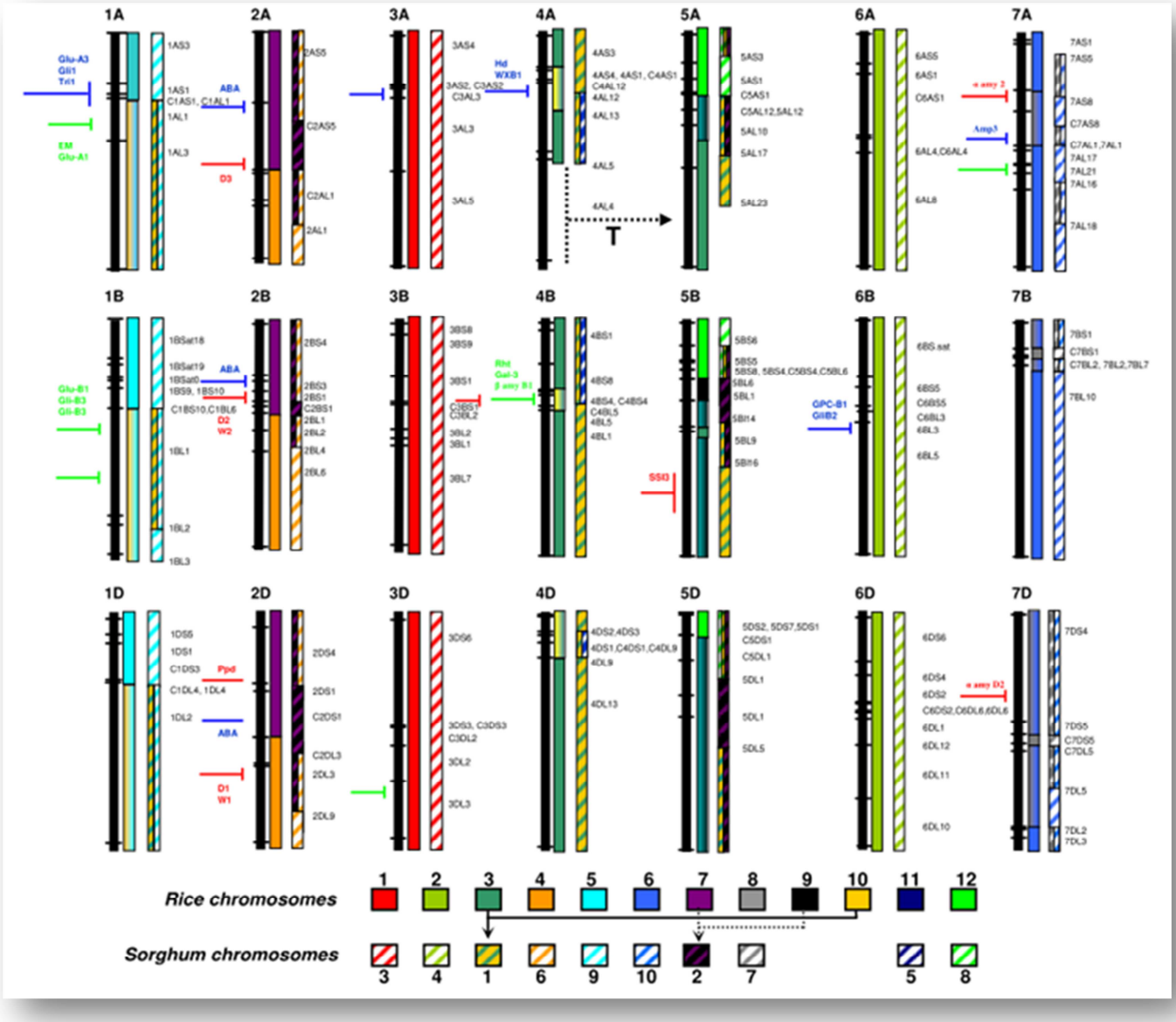


Figure 51. Visualisation de la diploïdisation phénotypique à travers 22 Meta-QTLs du blé tendre
Les 21 chromosomes du blé tendre sont représentés verticalement par leurs cartes génétiques ainsi que les relations de synténie avec le riz et le sorgho (*cf* code couleur, en bas). Les chromosomes du blé sont illustrés sur la base d'une carte consensus (barres noires verticales), les lignes noires horizontales représentant les bins de délétions (lignées de sears ; nomenclature à droite). Les 22 Meta-QTLs sont illustrés en rouge pour le rendement (8 MQTLs), bleu pour la teneur en protéines (8 MQTLs) et en vert pour la qualité boulangère (6 MQTLs). La synténie entre le blé et le riz, le sorgho est illustrée par des barres verticales à droite des chromosomes du blé, selon le code couleur des 12 chromosomes du riz et 10 de sorgho (indiqué au bas de la figure). Les gènes candidats contrôlant les méta-QTLs sont indiqués à côté des Meta-QTLs. *Source : Umar Masood Quraishi.*

Diploïdisation des caractères - Les résultats précédents relatifs aux caractères agronomiques majeurs et quantitatifs suggèrent que, malgré la redondance génique résultante de la polyploïdie (62% *cf*. chapitre I, page 41), les phénotypes ne semblent contrôlés, dans la grande majorité des cas, que par l'une des trois régions homéologues. Ainsi, cette spécialisation est reflétée par un seul allèle spécifique qui apporterait le gain (structural et fonctionnel) par rapport aux autres chromosomes portants les copies homéologues. D'après Quraishi *et al.* 2011 et Feldman *et al.* 2012, le blé moderne est caractérisé par une répartition asymétrique des caractères agronomiques portés par les trois génomes homéologues, ce qui définit ici la diploïdisation des caractères (ou des phénotypes au sens large). Dans la méta-analyse menée par Quraishi *et al*. les QTLs de teneur en protéine du grain sont majoritairement portés par le sous-génome A (*cf.* Figure 51). De la même manière Feldman *et al.* 2012 suggèrent que cette asymétrie phénotypique n'est pas aléatoire avec, parmi les 70 QTLs de domestication étudiés (rachis fragile, date d'épiaison, hauteur de la plante, la taille des grains et les composantes du rendement ; *cf.* Peng *et al.* 2003), un biais observé en faveur du sous-génome A. Les auteurs concluent que le sous-génome A du blé pourrait avoir joué un rôle plus important que le sous-génome B au cours de la domestication. La contribution des sous-génomes A et B a été étudiée par Feldman *et al.* 2012 et il apparaît, dans ce contexte, que le génome A contrôlerait préférentiellement les traits morphologiques y compris la structure de l'inflorescence, la forme des grains, les glumes non adhérentes (grain nu), la constitution générale de la plante, etc. Effectivement, ce sous-génome est porteur de nombreux gènes de domestication, tels que le gène Q (Zhang *et al.* 2011), codant pour un facteur de transcription APETALA2-like. Ce locus, 5A, apporte le caractère du libre-battage (épis tétraédrique avec un rachis non fragile). La pseudogénisation du locus Q-5B (mutations modifiant le cadre de lecture) ou la sous-fonctionnalisation du locus Q-5D (compartimentation de l'expression de ce gène dans différents tissus) contribuent également à ce phénotype de domestication mais dans une moindre mesure par rapport au locus 5A (Zhang *et al.* 2011).

A la suite des travaux menés dans les chapitres précédents sur l'asymétrie structurale et expressionnelle des sous-génomes du blé tendre, je souhaite aborder dans ce dernier volet, les liens entre plasticité génomique et phénotypique (c'est à dire la diploïdisation des caractères) en lien avec le phénomène de dominance des sous-génomes post-polyploïdie. L'hypothèse sous-jacente est que la dominance des sous-génomes, générée lors de la duplication totale de génome, peut dans une certaine mesure impacter la capacité de la plante à s'adapter aux contraintes de l'environnement. Les exemples précédents, montrent clairement une diploïdisation des phénotypes, que ce soit pour les gènes majeurs ou les caractères quantitatifs. Dans ce dernier volet, je propose d'étudier le phénomène de diploïdisation des caractères, au-delà des éléments bibliographiques précédemment présentés, *via* l'étude de l'inhibition du tallage chez le blé tendre porté par le compartiment paléo-sensible. Ce chapitre 3 est dans la continuité des deux précédents, permettant d'établir un continuum entre dominance des sous-génomes chez le blé tendre au niveau structural, expressionnel et phénotypique.

1.2. Etude du tallage chez le blé tendre

Le tallage chez le blé est la capacité de la plante à produire des épis ; c'est une composante forte du rendement. Donald en 1968, avait présenté l'idéotype de blé à fort rendement comme étant une plante monotalle à large épis (*cf*. Figure 52A). Depuis, quelques rares lignées monotalles ont été étudiées (Richards, 1988 ; Spielmeyer *et al.* 2004 ; Suenaga *et al.* 2005). Certains allèles permettent à la plante de

maintenir un rendement élevé, même sous certaines contraintes environnementales non favorables (Duggan *et al.* 2006).

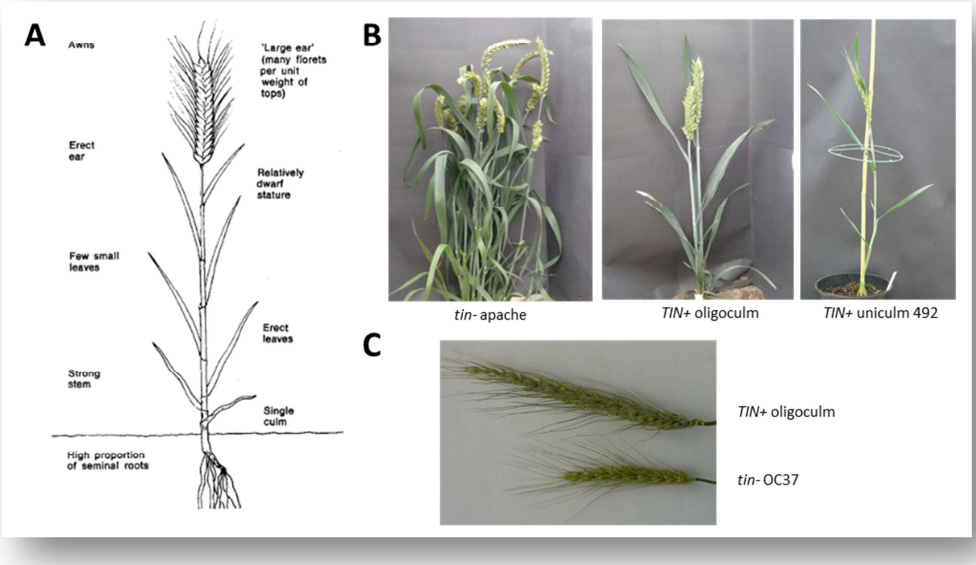


Figure 52. Idéotype et phénotype de divers blés
(A) Idéotype présenté par Donald en 1968. La figure illustre et décrit l'idéotype théorique de blé tendre performant pour le rendement ; monotalle avec de larges épis, une courte et forte tige, un puissant développement racinaire et quelques petites feuilles érigées. *Source : Donald 1968.* (B) Phénotype de divers lignées de blé tendre capable d'inhiber le tallage (*TIN*+) ou non (*tin*-). (C) Epis des 2 lignées contrastées croisées pour obtenir la population recombinante OxO ; 'Oligoculm' (*TIN*+) et 'OC37' (Orsay X Centurk lignée 37, *tin*-).

La capacité d'inhibition du tallage de certains blés semble relever d'un caractère majeur diploïdisé, puisque les premières études génétiques du tallage des blés (Richards, 1988) ont permis de détecter la présence d'un QTL majeur sur le bras court du chromosome 1A (et non 1B et 1D). Cet allèle, 1A, est le locus nommé TIN (pour Tiller Inhibition Number). Il est également associé au caractère de gigantisme des épis (*cf.* Figure 52B-C), apporté par la variété Israélienne 'Oligoculm' (Spielmeyer *et al.* 2004, Suenaga *et al.* 2005). Ce phénotype se caractérise par l'interruption de fonctionnement du bourgeon axillaire *via* l'internoeud basal, d'architecture solide plutôt que creuse. La croissance des bourgeons serait stoppée en raison du détournement de saccharose loin du bourgeon axillaire, pour soutenir l'allongement de l'internoeud (Kebrom *et al.* 2012). L'étude des capacités agronomiques de l'allèle TIN introgressé dans des cultivars (Duggan *et al.* 2006), semble révéler que le rendement du blé n'est pas diminué par la présence du locus TIN. En effet, certaines lignées TIN+ présentent moins de talles infertiles que les lignées TIN- (11 % en moins) et ont un nombre d'épillets par épi augmenté de 9 %, sans affecter la teneur en protéines du grain (Duggan *et al.* 2006). Ces capacités agronomiques pourraient s'expliquer par le fait que la plante produirait de plus grands épis pour compenser le nombre réduit de talles (Richards, 1988). Ce locus peut donc impacter très fortement le rendement et à ce jour, aucun gène n'a été caractérisé sur le chromosome 1A contrôlant le nombre de talles. Chez le blé, un mutant mono-talle a été décrit par Kuraparthy *et al.* 2007 impliquant un locus sur le chromosome 3 (mineur par rapport à l'effet du locus 1A), correspondant probablement au gène orthologue Uniculme4 (Cul4) identifié chez l'orge (Tavakol *et al.* 2015).

L'étude développée dans ce manuscrit, est basée sur la caractérisation du gène contrôlant le tallage (TIN), à partir du croisement entre les variétés 'Oligoculm' (TIN+) et 'OC37' (Orsay X Centurk lignée 37, TIN- ; *cf*. Figure 52C). L'approche de génomique translationnelle utilisée ici, permet de travailler ce type de caractère même si les espèces modèles diffèrent dans leurs architectures et physiologies. En raison de la disponibilité des ressources génomique, le riz et *Brachypodium* sont des espèces de choix pour transférer les connaissances et ressources chez le blé tendre. Ainsi, la génomique translationnelle permet de disséquer les bases génétiques du tallage et étudier la diploïdisation de ce caractère. Ce locus, positionné sur le chromosome 1A fait partie du groupe chromosomique 1 dit sensible, suite à la duplication ancestrale A1/A5, il y a 90 MYA (*cf*. Figure 53). La région paralogue du groupe chromosomique 3 est, elle, dite dominante. Ce locus est ainsi, un bon candidat pour approfondir l'impact de la dominance des sous-génomes post-polyploïdie induisant une compartimentation des caractères entre les trois sous-génomes, où le génome A contrôlerait préférentiellement les traits morphologiques (comme le tallage dans notre contexte) comme décrit par Feldman *et al.* 2012.

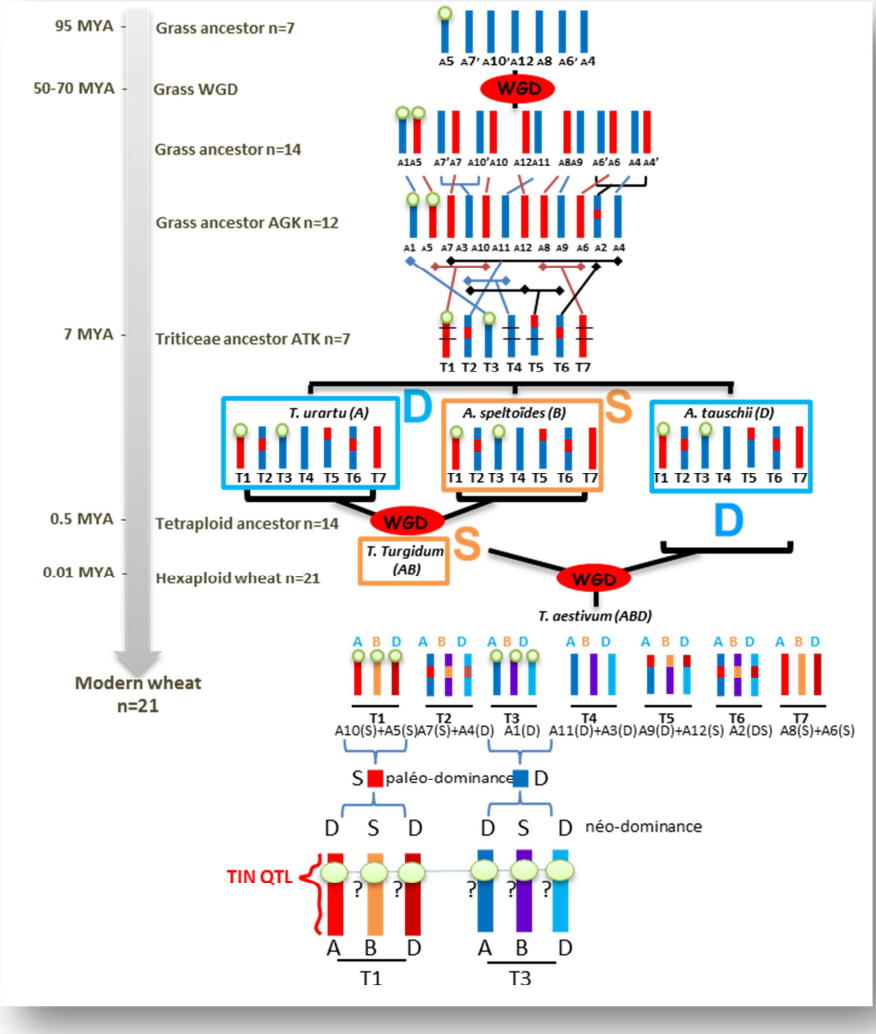


Figure 53. Régions paralogues/homéologues du locus TIN au regard de la dominance des sous-génomes du blé hexaploïde.

L'histoire évolutive du blé est schématisée à partir de l'ancêtre des céréales AGK à 7 chromosomes. Cet ancêtre a subi une duplication totale de son génome doublant le contenu chromosomique puis deux réarrangements pour aboutir à un ancêtre à 12 chromosomes. Cette WGD a engendré un effet de dominance des deux sous-génomes, matérialisés ici en bleu, pour la fraction dominante (A1-2-3-4-9-11) et en rouge, pour la fraction sensible (A5-6-7-8-10-12). Le locus TIN, matérialisé par un cercle vert et potentiellement présent en une copie dans l'ancêtre AGK, est en 6 copies chez le blé moderne. La région paralogue du locus est portée par le groupe chromosomique 1 sensible [T1(S)=A10(S)+A5(S)], et 3 dominant [T3(D)=A1(D)]. La néo-dominance des sous-génomes est matérialisée par des rectangles bleus sur les génomes diploïdes dominants A et D, et orange pour B sensible. AGK : ancestral grass karyotype ; ATK : ancestral *triticeae* karyotype ; WGD : whole genome duplication. La datation des évènements est indiquée en millions d'années (MYA).

2. Article

Physical mapping of the Tiller INhibitor (TIN) locus in bread wheat (*Triticum aestivum*).

Caroline Pont[1], Cécile Huneau[1], Maoine Elbaidouri[1], Florent Murat[1], Carole Confolent[1], Michael Throude[1], Mélanie Molinier[1], Florine Core[1], Annie Lebreton[1], Stéphane Bernard[1], Arnaud Bellec[2], Hélène Berges[2], Michel Bernard[1], Jérôme Salse[1].

[1]INRA/UBP UMR 1095 GDEC 'Génétique, Diversité et Ecophysiologie des Céréales'.
Laboratory PaleoEVO 'Paleogenomics & Evolution' (http://bit.ly/PaleoEvo) 5 chemin de Beaulieu, 63100 Clermont Ferrand, France. e-mail : caroline.pont@clermont.inra.fr Phone: +33(0)473624300 Fax : +33(0)473624453
[2]INRA, Centre National de Ressources Génomiques Végétales (CNRGV). Chemin de Borde Rouge BP 52627, 31326 Castanet Tolosan cedex, France.
Running title: Physical mapping of the Tiller INhibitor (TIN) gene in wheat (*Triticium aestivum*).

ABSTRACT.
- Tillering or vegetative branching is one of the most important component of shoot architecture in cereals as it contributes directly to grain yield and enhances plant phenotypic plasticity in response to environmental constraints.
- We report here the physical mapping and sequencing of the wheat tiller inhibition (TIN) locus unravelling a 109bp insertion in three independent reduced-tiller genotypes.
- The current results and resources open the way in monitoring tiller number in current breeding practices for improved tolerance to constraints.

INTRODUCTION.
Climate change, steady population growth and increasing demand from emerging economies are expected to threaten food security over the world in the following decades. With the increasing worldwide demand for food and unavailability of new land for cultivation, the new challenge for plant breeders is now to develop new plant ideotypes, which is referenced as ideal plant architecture (IPA). IPAs was described by Donald (1968) as high-yielding and durably stress-tolerant crops and more importantly with stable production in unstable environmental conditions associated with high temperature as precipitation fluctuations in the frame of the current and future climate changes.
Bread wheat (*Triticum aestivum*) is the third most-produced cereal worldwide with 713 million tons in 2015, after maize (1,016 million tons) and rice (745 million tons). In the context of producing high-yielding wheat varieties, Donald *et al.* (1968) described a wheat ideotype featuring a single strong

semidwarf culm, a large spike, and erect leaves. Atsmon and Jacobs (1977) selected wheat lines from a North African wheat cultivar and its derivatives, which, in part, resembled Donald's ideotype. These genotypes showed either a restricted number of tillers (Oligoculm) with extremely large (Gigas) spikes and leaves. The CIMMYT (International Maize and Wheat Improvement Center, *www.cimmyt.org)* exploited germplasm collections to develop new Gigas wheat types with larger sink capacities, a large number of grains per spike, and a large grain size potential (Rajaram and van Ginkel 1996). Richards *et al.* (1988) performed genetic analyses using the Gigas lines developed by Atsmon and Jacobs (1977) and demonstrated that the tiller inhibition was controlled by a single locus (TIN), which was physically linked to the gene controling glume pubescence (Hg) on chromosome 1AS. The authors also reported an association between tiller inhibition (Oligoculm) and large spike size (Gigas), which may be required to compensate for the restricted tiller number (Richards *et al.* 1988). The inhibition of tillering can cause major effect on entire plants with a significant influence of the genetic background and the environment. The root-to-shoot ratio increased with the TIN gene and most strongly at low temperatures (Atsmon *et al.* 1986, Hendriks *et al.* 2016). A separate de-tillering study confirmed greater root-to-shoot ratios with regular tiller removal in non-tin-containing genotypes. The tiller number has slowed and delayed water use, while maintaining yield potential (grain number) resulting in greater water availability, greater stomatal conductance, cooler canopy temperatures, and maintenance of green leaf area during grain-filling (Hendriks *et al.* 2016). As suggested by Atsmon in 1986, the Gigas trait appears to derive from the capacity to use the savings from restricted tillering for both greater leaf and ear growth per shoot and allows the plant to resist at some harsh environmental conditions (Mitchell *et al.* 2013), overall the TIN lines producing yield comparable (grain number and kernel weight) to free-tillering lines in terminal water stress environments.

Tillering or vegetative branching is one of the most important components of shoot architecture in cereals as it contributes directly to grain yield (Kebrom *et al.* 2013) and it has been reported as a trait enhancing plant phenotypic plasticity in response to environmental constraints (Ihsan *et al.* 2016). In grasses, tillers are lateral branches (*i.e.* culms) that grow from nodes of unelongated internodes at the base of the plant, affecting important agronomical features such as competition with weeds and ease of harvesting. Primary tillers arising from the main culm initiate new axillary buds that may, in turn, develop into secondary tillers and so on in a reiterative pattern. In rice the IPA traits include low tiller numbers with few unproductive tillers, more grains per panicle than the currently cultivated varieties (Jiao *et al.* 2010). The *IPA1* (Ideal Plant Architecture 1) which profoundly changes rice plant architecture and substantially increases grain production encodes OsSPL14 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 14) and is regulated by a microRNA (OsmiR156, Jiao *et al.* 2010).

Evolutionarily conserved genetic pathways controlling axillary branching in both monocots and eudicots have been reported (Kebrom *et al.* 2013; Janssen *et al.* 2014; Waldie *et al.* 2014; Li *et al*. 2003) in involving major genes such as GRAS (for GIBBERELLIC ACID-INSENSITIVE, REPRESSOR of GA1, SCARECROW and MONOCULM1), LATERAL SUPPRESSOR (Raatz *et al*. 2011), Uniculm encoding an ANKYRIN protein (Tavakol *et al.* 2015).

In wheat, despite some recent progress (Dabbert et al., 2010; Mascher et al., 2014), few genes controlling tillering number has been dissected. Kuraparthy *et al.* (2007) described, based on wheat tillering mutant producing one main culm, the tiller inhibition number (tin3) locus located on chromosome 3A at orthologous position to Uniculme4 (Cul4) in barley (Tavakol *et al.* 2015). Using

classical molecular markers and doubled-haploid (DH) population resulting from a cross between a Japanese cultivar 'Fukuhokomugi' and an Israeli wheat line 'Oligoculm', Suenaga *et al.* (2005) unraveled a major QTL on the wheat chromosome 1AS (TIN1A). This QTL was stable in both field and greenhouse conditions in controlling spike number per plant and was close to the glume pubescence locus (Hg). Secondary loci with segregation distortion were clustered on chromosomes 4B, 4D, 5A, 6A, 6B, and 6D, making TIN1A as the unique major locus (*i.e.* gene) driving the number of tillers in bread wheat.
Here, we report the physical mapping of a major QTL related to tiller number identified on the short arm of chromosome 1A deletion BIN 1AS3-0.86-1.00, as reported in the previous study (Spielmeyer *et al.* 2004), based on a translational research approach between wheat and the grass relatives.

RESULTS.

TIN locus genetic mapping.

TIN (number of tiller) and Gigas (spike size) traits were dissected using 423 wheat lines from a F6 SSD population obtained from a cross between Oligoculm and a french breeding line OC37 (Orsay X Centurk n° 37), referenced as OxO population. From a subset of 186 lines maximizing the recombination pattern observed in the OxO population, a genetic map was constructed from 230 SSR and 29 Conserved Orthologous Set (COS targeting the chromosome 1A) markers selected from a Wheat Consensus Genetic Map (WCGM2013, Pont *et al.* 2013) consisting in 7,520 molecular markers covering 4,318.03 cM with a marker density of one marker every 0.78 cM and including mapped RFLP (1,687), AFLP (712), STS (262), SSR (2,315), DaRTs (1,246) and COS-SNP (375) markers. QTL detection allowed us to identify 20 QTL with a LOD score >3 and r2 >5% (*cf.* supplementary data S1) with a major QTL for tiller number located on the short arm of chromosome 1A (deletion BIN 1AS3-0.86-1.00) with a LOD score of 22.4 and explaining 49% (r2) of the observed phenotypic variation (Supplementary Figure 1). The interval co-located to the TIN interval as reported initially by Spielmeyer *et al.* 2004, Figure 1A.
QTL fine mapping was performed based on the wheat syntenome that consists of 72 900 (73.4 % of the 99 386 gene models available from IWGSC 2014) genes ordered on the 21 chromosomes (Pont *et al.* 2013, ElBaidouri *et al.* 2016). The wheat syntenome of the chromosome 1A consists in 1121 ordered genes with 76 COS-SNP markers mapped on the wheat consensus map WCGM2013 (Pont *et al.* 2013). Mapping in the OxO population of the 76 COS-SNP markers deriving from the synteny between the wheat chromosome 1A and chromosomes 5, 9 and 2 of respectively, rice, sorghum and *Brachypodium* genomes refined the QTL interval to 0.96 cM (interval mapping with likelihood of 95%; Supplementary Figure 1 and Supplementary Table 1). Three additional QTLs have been detected on the same locus dealing with spike length (r2=29% in 1.32cM), flower number per spikelet (r2=29% in 1.74cM), and spike compactness (r2=11.4% in 2.49cM). The QTL interval (Figure 1A-A) is located between COS257 (corresponding to LOC_Os05g01310, Bradi2g39890, Sb09g000390 genes respectively in rice, *Brachypodium* and sorghum) and COS229 (corresponding to LOC_Os05g01290, Bradi2g39930, Sb09g000370 genes respectively in rice, *Brachypodium* and sorghum) (*cf.* figure 1a). The COS257 (Figure 1A-B) marker is close to gwm136 SSR marker (0.3cM), previously reported to be closely associated to the TIN gene (Kumar et al. 2015; Spielmeyer *et al.* 2004).

TIN locus physical mapping

We used the previous COS markers to screen wheat BAC libraries from *T. aestivum* cultivar Renan, *T. aestivum* cultivar Chinese Spring and *T. aestivum* cultivar Oligoculm (*cnrgv.toulouse.inra.fr*). This libraries are organized in 384 wells plates of BAC pools with a specific screening strategy to reduce the number of PCR reactions in order to identify a single clone of interest using agarose gels (*cf.* Materials and Methods). Positive pools were analyzed by SSCP or real-time PCR to assign wheat chromosomes using nullisomic-tetrasomic lines. By crossing the pool coordinates (plates/lines/columns), we identified the BAC clones harboring the 1A TIN locus. Using this strategy, 21 BAC clones covering the entire TIN locus have been identified. In order to complete the DNA sequence of the TIN interval, we also used 5 BAC clones from the *T. aestivum* cultivar 'Chinese Spring' MTP1AS (Breen *et al.* 2013). The individual BAC clones of interest were sequenced on the GS Junior (Roche) using the 454 and MIseq illumina sequencing technologies at the genotyping platform GENTYANE (UMR-1095 GDEC, Clermont-Ferrand), according to the procedure described by the manufacturer. These 26 BACs were assembled and assigned to the subgenomes A (9 from oligoculm, 8 from Chinese Spring, 4 from Renan), B s (1 from oligoculm, 2 from Chinese Spring, 1 from Renan) and D (1 from oligoculm). Finally, the 21 BAC clones from the A subgenome were ordered using the wheat chromosome 1A pseudomolecule from *T. aestivum* cultivar Chinese Spring (NRgene draft genome sequence, https://urgi.versailles.inra.fr/download/iwgsc/IWGSC-WGA/).

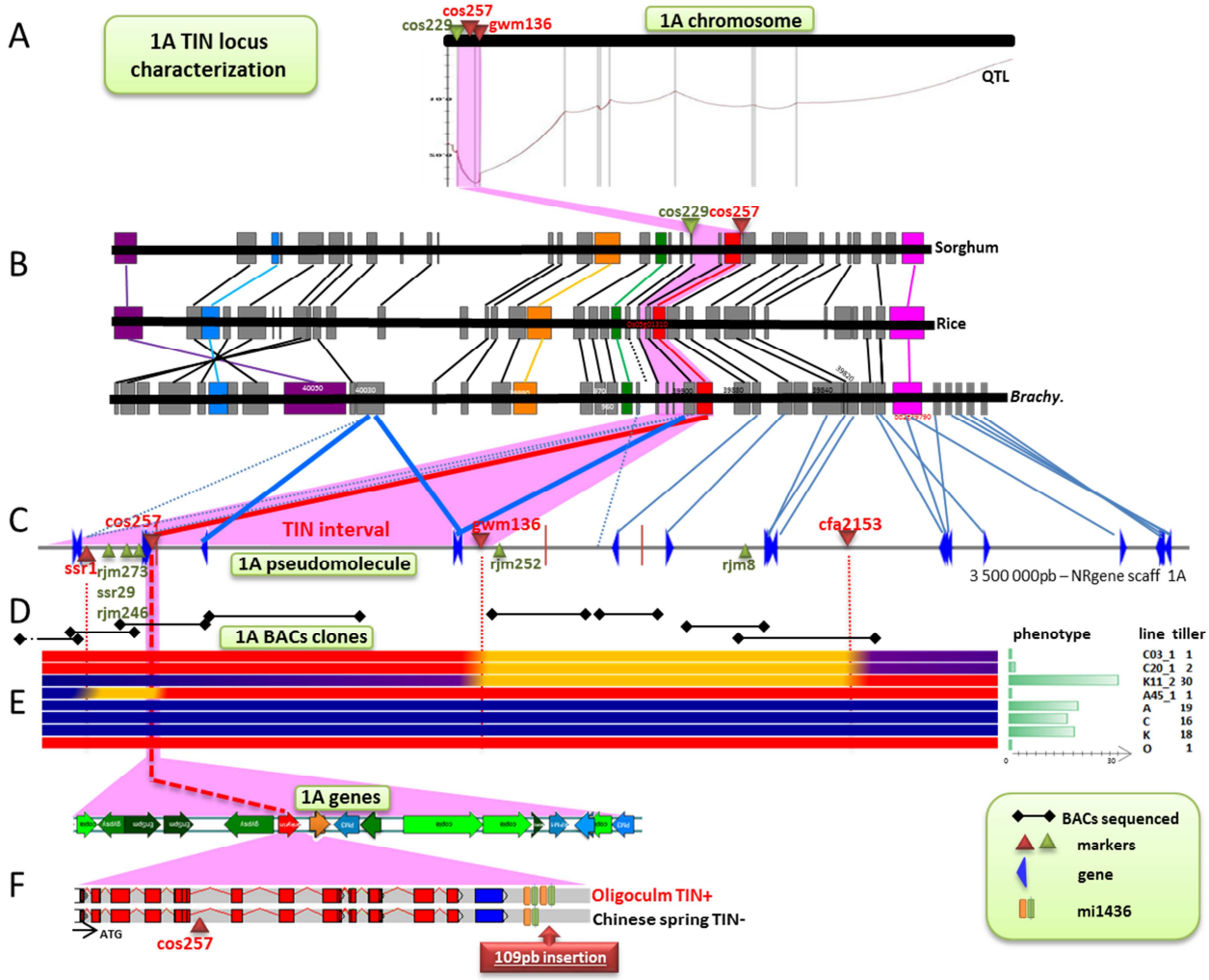


***Figure 1*. Synteny-based characterization of the TIN locus in bread wheat.** *(A) QTL detection (LOD score) from the Oligoculm x OC37 population on the genetic map of wheat chromosome 1AS. The significant interval (0.4 cM) is obtained between 2 markers gwm136 and COS229. (B) Graphical representation of the orthologous region in sorghum (Chr9), rice (Chr5), and Brachypodium (Chr2) within conserved genes linked with black lines. The TIN interval is defined with COS markers COS257 (corresponding to rice / Brachypodium / sorghum gene respectively, LOC_Os05g01310 / Bradi2g39890 / Sb09g000390) and COS229 (LOC_Os05g01470 / Bradi2g39790 / Sb09g000490) highlighted in pink. (C) The TIN interval from the 1A Pseudomolecule from the NRgene draft genome sequence is annotated with genes in blue. Sequenced based RJM and SSR markers are positioned on the 1A sequence interval. (D) Anchored and annotated 1AS BAC clones in the TIN interval are illustrated by horizontal black lines. (E) Recombinant (and parental) lines with their phenotype (spike number) are illustrated for QTL fine mapping. (F) Micro-synteny of the TIN region with the ankyrin repeat family protein (red box) and after the 3'UTR (in blue) an insertion of 109pb in Oligoculm corresponding to mi1436.*

TIN locus sequencing and annotation

The TIN locus is covered by a 3.5 Gb region from the chromosome 1A pseudomolecule covering the assembled BAC clones (Figure 1A-C). Based on the TIN locus sequences we are able to derive new sequence-specific markers (Supplementary Table 1) in order to fine map the QTL based on BC2F3 populations obtained from a cross between Oligoculm and three elite lines (Koreli, Apache, Caphorn). Complementary in recombination pattern at the TIN locus in this material showing less than 20% (and distinct pattern between the three populations) of Oligoculm introgression outside the locus of interest (Supplementary Figure 2) allowed us to correlate phenotyping and genotyping data within an interval of 933 600 kb and involving 18 genes (*cf* Figure 1A-C). Genes and repeated elements were annotated (*cf* material and method section) and taking advantage of the presence of unique TE insertions junctions in the genome, we was able to design specific primers. We used 'RJPrimers' tool reported by You *et al.* (2010). We designed 129 RJPrimers to anchor BACs and to order the physical map using agarose gel to reveal the presence/absence polymorphism of such TE junctions in Renan, Chinese spring and Oligoculm BACs. In the same way we design 32 SSR primers using the SSRfinder tool developed by Steven Schroeder (Sharopova N *et al.* 2002). 31 RJM and 13 SSR were precisely identified as 1AS-specific and mapped on the chromosome using deletion lines. These markers were used to screen the recombinant lines (425 plants F3) from crosses between 3 cultivated bread wheat lines 'Apache', 'Caphorn' and 'Koreli' (TIN-phenotype), and 7 lines from the mapping population 'OxO' (TIN- phenotype). The three Backcrosses maximize the opportunity to find polymorphism and recombinant lines within the TIN interval. This strategy yielded 2 recombinants lines between gwm136/cfa2153 from OxCaphorn (lines C20_1 and C03_1), one between SSR1/COS257 from OxApache (line A45_1) and one between gwm136/cfa2153 from OxKoreli (K11_2), Figure 1A-E.

Within the reduced TIN interval (COS257/gwm136), 3 orthologous genes are found corresponding to Bradi2g39890, Bradi2g39900, Bradi2g40030 (Figure 1A, broad lines blue/red) and the Pm3 like gene already published (Wicker *et al.* 2007, Breen *et al.* 2013). Bradi2g39890 (red broad lines) is an ankyrin repeat family protein TPR10 (tetratricopeptide repeat 10) encoding one of the 36 carboxylate clamp (CC)-tetratricopeptide repeat (TPR) proteins (Prasad *et al.* 2010) that can potentially interact with a

Hsp90/Hsp70 as co-chaperones. Bradi2g39900 is a copper ion binding (helix domain-containing protein) and Bradi2g40030 is a Phosphatidylinositol 3-kinase catalytic subunit type 3 and is repeated twice on the region. This latest gene is not found on the Chinese spring sequence and only few exons are found on Renan and Oligoculm BACs sequence, suggesting a pseudogene structure.

TIN locus candidate sequence

We re-sequence the TIN region of the two main conserved genes in several wheat genotypes. The comparison of the amino acid sequence of the annotated ankyrin gene (Figure 1A-F, red line) with public database shows that this gene is highly conserved (see supplementary data S4) but with sequence divergence at the 5'UTR and 3' UTR. We re-sequence 3' UTR region with the primers F24 (CATTGCCAGCATACATTCTC) and R23 (GCTGACACGGGTTTTTAT) in several TIN+ genotypes: Oligoculm (material of the current study); Pubing3558 (Zhang *et al.* 2013); Bank + tin (Richards *et al.* 1988) and uniculm (Richards *et al.* 1988) (see Materials and Methods). All TIN+ lines show deletion / mutations (Figure 2) downstream the taTPR 3' UTR (Figure 1, bottom). None of the other regions from the TIN locus that have been sequenced show a clear discrimination at the amino acid level between TIN+ and TIN– genotypes (Figure 2). Overall, Oligoculm, Bank +tin, Pubing3558 and uniculm (all TIN+) lines show the same 109 pb insertion encoding a miRNA1436 conserved in rice (Sunkar *et al.* 2008) and barley (Kantar *et al.* 2010). This micro-RNAs was also described in *Festuca arundinacea* (Unver *et al.* 2010) and in wheat as a highly represented repeat-related miRNA family in 1AL survey sequences (Lucas *et al*. 2012). Based on this insertion it was possible to genotype all recombinant lines based on a perfect F24R23 diagnostic primer pair (with TIN+ lines showing the AA_{F24R23} allele, 109pb insertion). This marker showed cosegregation with the TIN phenotype for all the recombinants identified in the targeted interval.

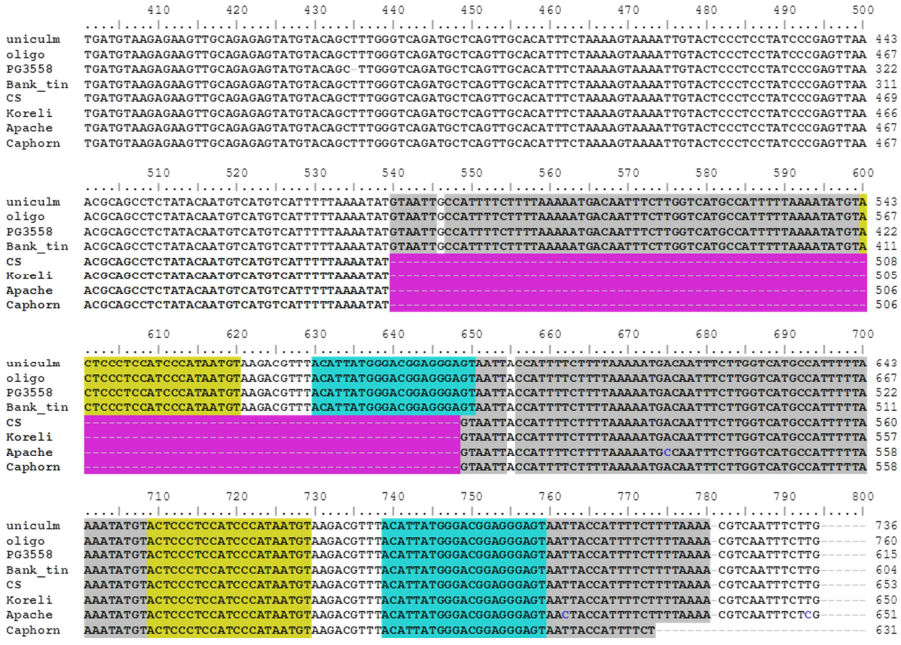


Figure 2. A 109 pb insertion as candidate for the TIN trait. *ClustalW of the sequenced 3' UTR region with the primers F24 (CATTGCCAGCATACATTCTC) and R23 (GCTGACACGGGTTTTTAT) in several TIN+ lines*

(Uniculm, Oligoculm,PB3558, F4-11 line, Bank + tin) and TIN- lines (Chinese Spring, Koreli, Apache and Caphorn). TIN+ lines show a 109 pb insertion (deletion highlighted in pink) corresponding to the miRNA1436. The miR1436 mature is highlighted in blue and the complementary one is highlighted in yellow. The upstream region is repeated and highlighted in gray. Regions with SNPs between TIN- lines are highlighted in blue.

Using the wheat microRNA portal (http://wheat.bioinfo.uqam.ca./index.php) we search experimental conditions were the miRNA1436 were expressed. This specific 21 nt ACAUUAUGGGACGGAGGGAGU was found in abundance in leaves but not in spikelet in Chinese spring cultivar. We also identify by blastn (see methods) the presence of this specific miRNA in drought-susceptible cultivar Zhengyin1 at the two-leaf stage from the study of Ma *et al.* 2015 but not in Hanxuan10 which is a drought-tolerant cultivar. This miRNA (named 'wheat-miR-304' in Ma *et al.* 2015) shows a differential expression only in the drought-susceptible wheat genotype after a dehydration stress subjected at the two-leaf stage seedlings. We investigated the expression (QPCR) pattern of miRNA1436 (using the surrounding genes as control for Bradi2g39930, Bradi2g39890 and Bradi2g39840) during the early developmental stages in Oligoculm and the recombinant lines. The three control genes appear to be expressed in both TIN+ and TIN- lines. Regarding miR1436, the expression was observed by the smallRNA-seq approach from the both the recombinant line Apache A45 TIN- (primers BB_{F24R23}) and Oligoculm TIN+ (primers AA_{F24R23}), see Materials and Methods. In order to complete the expression analysis we performed the miRNA1436 expression analysis for the whole plant (pooled root-stem-leaf tissues) at two developmental stages for Oligoculm and from the TIN- lines Apache A18 (see Supplementary Figure 3). miRNA1436 expression (expressed in RPKM) increased for the TIN- (primers BB_{F24R23}) line between 2 tillers and 8 leafs stages whereas miRNA1436 does not show different in expression for Oligoculm (TIN+, see Supplementary Figure 3). This data supports a difference in expression for miRNA1436 before and after the tillering stage. Hendriks *et al.* 2016 reported similar phenotypes between TIN + (bank +tin) and TIN- (bank) until tillering where differences appear with TIN + showing reduced biomass of up to 60% in total leaf area and in contrast increases in total root length of up to 120% and root biomass up to 145%.

DISCUSSION

We report here the physical mapping of the Tiller Inhibitor (TIN) locus where for Oligoculm and three low tillering lines (Bank +tin, uniculm and Pubing3558) we identified a 109pb insertion encoding a miRNA1436 after the 3'UTR of a wheat gene ankyrin repeat family protein (ta-TPR10). The wheat line '492' (uniculm donor of the TIN gene in 'Bank +tin', see methods) and the wheat line '380' (donor of TIN gene in Oligoculm) have been also proposed to carry the same TIN gene (Spielmeyer *et al.* 2004). From a doubled haploid (DH) population of 110 lines from a cross between the Japanese winter wheat 'Fukuhokomugi' and the line 380, Suenaga *et al.* 2005 concluded that the lines 492 and 380 carry the same TIN gene. The line 380 probably contained additional modifier genes that produced more tillers than the line 492. *Agropyron cristatum* (L.) Gaertn. (2n=4x=28 genomes PPPP donor of TIN gene in 'Pubing3558' line, (see Materials and Methods) is a fairway crested wheatgrass introduced in the arid areas of the western United States for forage production from Asia. It possess many desirable traits, such as high tiller number, high floret numbers, and resistance to wheat rusts, powdery mildew, and barley

yellow dwarf virus, and the introgression of chromosomal segments into common wheat could enhance thousand-grain weight and spike length (Zhang *et al.* 2015).

The 109pb insertion match with Miniature Inverted Terminal repeat Elements (MITEs) from the Stowaway family, Mariner subfamily CACTA and encode the miRNA1436 as referenced in Lucas *et al.* 2012. This miRNA is monocot-specific (Schreiber AW et al. 2011), first identified in 20 loci in rice from drought-stressed sequenced library (Sunkar *et al.* 2008) and then in barley (Kantar et al. 2010). This microRNA was also identifying in *Festuca arundinacea* in response to foliar glyphosate with miRNA1436 down regulated upon in response to (5 and 20%) glyphosate treatment compared to control samples (Unver *et al.* 2010).

Similar to rice and barley, the wheat miRNA1436 was found expressed in leaves and in response to abiotic stress. This 21nt miRNA, named 'wheat-miR-304' in Ma *et al.* 2015, shows a differential expression only in the Zhengyin1 drought-susceptible wheat genotype after dehydration stress applied at the two-leaf stage. In Tang *et al.* 2012, miRNA1436 were identified in their data sets from the thermosensitive genic male sterile (TGMS) bread wheat lines hypersensitive to low temperature during meiosis. Kurtoglu *et al.* 2014 studied wheat microRNA using whole-genome sequence and identified miR1436 as well as its potential target on chromosome 5D, and few in rice (Sunkar *et al.*, 2008). Using psRNATarget tool (http://plantgrn.noble.org/psRNATarget/) with the selected preloaded transcript library for target search *Triticum aestivum*, unigene, DFCI Gene Index (TAGI), version 12, released on 2010_04_18 it was possible to identify 59 putative targets for miR1436 in wheat (Supplementary Table 2). In Oligoculm the TIN+ allele (*i.e.* 109 bp insertion) makes the miRNA1436 with a dicistronic structure compared to TIN- alleles (supplementary Figure 4). In plants, the majority of the miRNA families are conserved and clustered (Merchan et al. 2009) but the dicistronic miRNA1436 (TIN+ allele) is monocot-specific (Schreiber AW et al. 2011) and in wheat the miRNA1436 precursor is specific from the 1AS TIN locus explaining here that the major TIN locus has been identified on the chromosomes 1AS.

MATERIALS AND METHODS.

Plant material - The mapping population 'OxO' (Oligoculm x (Orsay X Centurk n° 37)) was sown in autumn 2006 and harvested in summer 2007 (423 F6 SSD lines), in Clermont-Ferrand (France). To construct a genetic map with COS markers and SSR markers, 186 lines were selected. The uniculm wheat line '492' is originated from the the Weizmann Institute of Science laboratory (Israel) by selections for five generations within a heterogeneous population which was introduced as a North African local cultivar (Atsmon *et al.* 1977). The wheat line '492' and the Oligoculm wheat line '380' are full sibs and originated from a cross between the cultivar 'Alpha' and a North African land race (Atsmon *et al.* 1986). A near-isogenic line referred to as 'Banks + tin' was produced by backcrossing (BC 4) the tin from a progenitor line of 492 into the Australian cultivar *Triticum aestivum* 'Banks' (Richards *et al.* 1988). The F4-11 and F4-33 lines come from a single heterozygote plant create by crossing RWG18 and ND495 (RWG18 was made from a Langdon-*T. dicoccoides* chromosome 2A disomic substitution line and AL8/78 and ND495 is a spring type genotype). Pubing3558 comes from a wild grass *Agropyron cristatum* x Fukuhokomugi (Zhang *et al.* 2013). Lines were obtained from the Biological Resources Centre (INRA GDEC1095 http://www6.clermont.inra.fr/umr1095/Centre-de-Ressources-Biologiques).

About 3 g of tissue was ground in liquid nitrogen and 1 g were extracted in 4.5 ml of buffer (0.1 M NaCl; 10 mM Tris HCl pH 7.4; 1 mM EDTA pH8; 1% SDS) with 3 ml phenol–chloroform–IAA (25 :24 :1). After

mixing and centrifugation, the supernatant was extracted twice with 3mlphenol–chloroform–IAA (25:24:1). After mixing and centrifugation, the supernatant was extracted twice with 3mlphenol–chloroform–IAA (25:24:1). The aqueous phase was precipitated by addition of 300µl sodium acetate (3 M pH 5.2) and 6 ml of ethanol (100%). Then DNA was purified by ethanol precipitation, quantified and adjusted at 10 ng/µl.

Genotyping procedure - *Conserved orthologous set (COS)* - Wheat EST-contigs exon structures were identified through rice/wheat sequence alignments, as conserved HSPs correspond to exons (Quraishi *et al.* 2009). The precise position of the Exon-Intron junction is provided to Primer 3 package to select primer pairs on exons for intron amplification with the following parameters suitable for detection on Applied Biosystems (ABI) capillary sequencer: (1) Primer size (20 to 25 mer as default parameters), (2) Amplicon size (between 250–800 bp as default parameters), (3) Tm (between 57–63 as default parameters), (4) GC clamp (equal to 2, i.e. a G or C at the 5' extremity as default parameters), (5) GC percentage (50% as default parameters). ***Single Sequence Repeat (SSR)* -** Microsatellite markers were selected from the consensus genetic map WCGM2013 (Quraishi *et al.* 2011, Pont *et al.* 2013). PCR was carried out for SSR markers using M13 protocol (Nicot *et al.* 2004). Briefly the DNA amplicons were analyzed in a protocol using Applied Biosystems® FAM Fluorochrome using 384-well plates. PCR (15 µl per well) with 30 ng of DNA were performed with AmpliTaq Gold® PCR Master Mix (Applied Biosystems) according to the protocol. The standard PCR program for amplifying DNA was a denaturation step at 95°C for 5 min, 35 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s; and a final extension at 72°C for 5 min. The amplifications were performed on ABI PRISM 3130xl (Applied Bio-system, Foster City, USA) on non-denaturing conditions. The size of fragments and polymorphism analysis were performed using GeneMapper v4 software (Applied Bio-system). ***Single Strands Conformational Polymorphism (SSCP)*** - COS primers were synthesized with 5' extensions in order to facilitate the labeling procedure at low cost: forward primer with the CACGACGTTGTAAAACGAC sequence extension and reverse primer with the CAGGAAACAGCTAT GACC sequence extension. Polymerase chain reaction (PCR) fragments were produced in two steps. In a total volume of 15µl, genomic DNA (30 ng) was first amplified with the following PCR mix: 10 mM Tris–HCL, 3.1 mM MgCl2, 50 mM KCl, 0.001% gelatine pH 8.3, 5% glycerol, 400 µM dNTP, 0.4 µM forward and reverse primers, and 0.2 U Taq polymerase (Qiagen). This PCR product was diluted (1/10) and reamplified with the same PCR mix including 0.2 µM of each labeled primers (6-FAM and NED, Applied Biosystems) in a final volume of 15µl. Two microliters of the PCR product was then diluted (1/10) and pooled with 0.2µlof 900 bp MegaBase ET900-R Size Standard (GE Healthcare), 0.2µl of 0.3 N NaOH and 9µl HI-Di Formamide (Applied Biosystems). Fragments were separated by capillary electrophoresis on ABI3100 (Applied Biosystems) in 50 min with a 36 cm capillary. The running polymer consists in 1× of running buffer, 5% Genscan Polymer (Applied Biosystems), and 10% glycerol. Samples were denatured during 2 min at 95 C and 10 min in ice. The sample buffer consists in 1× of running buffer and 10% glycerol. After denaturing, the samples were injected at 2.5 kV during 50 s and separated at 18 C, 25 C, and 35 C at 15 kV. Data were analyzed using GeneMapper 3.7 software. ***Kaspar*** - The KASPar (KBioscience Ltd., Hoddesdon, UK) assay was used to stufy SNPs and homeoSNPs. All assay primer sets were designed by KBioscience (KASpar-By-Design) and assay screening and genotyping were performed on the LightCycler® 480 Real-Time PCR System (Roche Applied Science) with KASPar SNP reagent Mix and 2.5 ng of genomic DNA. Details of the method used can be found at

http://www.kbioscience.co.uk/. Analysis was performed using LightCycler 480 software with end point genotyping module.

Phenotyping and QTL detection - The plants were phenotyped in two replicats for different traits evaluation: date of anthesis, length spike, spike number per plant, spikelet number per spike, flower number per spikelet, Hairy glume, plant length, TKW (thousand kernel weight).Genetic map of Oligoculm x OC37 were made using MapMaker Version 3.0 (Lincoln *et al.* 1993). Centimorgan (cM) values of recombination were calculated using Haldane mapping function of the software. In this program two-points/group command was used for establishing possible linkage groups with minimum LOD of 3.0 and a recombination fraction of 0.37. QTL were calculate from the map generated from mapmaker using software MultiQTL version 2.5. MultiQTL software integrates a broad spectrum of data mining, statistical analysis, interactive visualization and modeling tools, that allow: single-QTL, two-linked QTLs and multiple QTL analysis (marker and interval); one-, two-, and multiple-trait analysis. The software evaluates the accuracy of the estimated model parameters (QTL effect and its chromosome position) by significance testing and comparing of alternative models by using permutation, and a Monte-Carlo analysis. A QTL was considered significant when its LOD score was superior or equal to 3 and heritability greater than 5%. Confidence intervals of QTL parameters (location, effect…) were estimated using a 1,000-bootstrap resampling approach.

BAC clone screening and sequencing

BACs library from T. aestivum cv. Chinese Spring, "Tae-B-Chinese spring" and cv T. aestivum 'Renan' [INRA – CNRGV [http://cnrgv.toulouse.inra.fr/] was used in this study. The Chinese Spring library (9.0x genome coverage) consists of 1147776 clones organized in 2988 384-well plates. The BAC library from 'Renan' (6.9x genome coverage) consists of 812 544 clones organized in 2592 384-wells plates. For each 384-wells plate, 16 line pools and 24 column pools and a total of 40 pools are available. Overall, 274 line and column pools are available to identify BAC clones of interest through PCR screening. On agarose gel each amplicon (or BAC of interest) in a 384-wells plate is identified by coordinates which identify positive line and column pools in a matrix. By crossing the coordinates of positive pools, we determined the plate harboring the clone of interest. The individual BAC clone of interest is then identified on the specific plates by PCR screening. BAC sequencing was performed on the GS Junior (Roche) using the 454 sequencing technology and miSEQ (illunima) according the procedure described by the manufacturer at the genotyping platform GENTYANE (UMR-1095 GDEC, Clermont-Ferrand).

BACs clone from T. aestivum cv. Olicogulm, was screened in this study

We screen also non gridded BAC library from TIN+ cultivar 'Oligoculm' construct by CNRGV platform [INRA – CNRGV [http://cnrgv.toulouse.inra.fr/]. They offers the possibility to construct a targeted BAC non gridded library directly from the genotype of interest, starting with a 1x genome coverage (with mean inserts size around 110kb). Pools are screened using real-time PCR or on agarose gel, and the BAC clones carrying the markers are then individualized and sequenced.

BAC clone annotation - Genes and repeated elements (TEs) were identified by computing and integrating results based on BLAST algorithms, predictor programs and software described below. Gene structure and putative functions were identified by combining results of BLASTN and BLASTX alignments

against dbEST (NCBI) and SwissProt databases with the results of a gene predictor program, FgeneSH [60] with default parameters. Known genes were named based on BLASTX results against protein with known functions (SwissProt). Transposable elements (TEs) were detected by blast-based comparison with two databases of repetitive elements: TREP (Graingenes) and Repbase (Jurka J *et al.* 2000). TEs boundaries were identified through Repet package (Flutre *et al.* 2011). Insertion profile of TEs is identified using TEnest, a modified version of svg_ltr.pl script (Kronmiller *et al.* 2008).

miRNA expression analysis - RNA was extracted and purified according to Pont et al. 2011. The whole plant was ground in liquid nitrogen and RNA extracted with tris buffer and phenol–chloroform. 3µg of RNA were sent to integragen institute (http://www.integragen.com/fr/) and samples were sequenced on HIseq illumina technology according to the procedure described by the manufacturer to obtain 20 million of reads per sample. Mapping reads were performed with Bowtie software (Langmead et al. 2012) and visualize with IGV (Integrative Genomics Viewer; http://software.broadinstitute.org/igv/).

SUPPLEMENTARY DATA (*cf*. Annexes du manuscrit)

REFERENCES

Atsmon MG Bush MG Bush LT Evans LT Evans. Effects of Environmental Conditions on Expression of the 'Gigas' Characters in Wheat. Functional Plant Biology (Impact Factor: 3.15). 01/1986; 13(3):365-379.

Lincoln, S. E., M. J. Daly, and E. Lander. 1993. Mapmaker/EXP 3.0 Manual. Whitehead institute for biomedical research.

Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J. The 'inner circle' of the cereal genomes. Curr Opin Plant Biol. 2009 Apr;12(2):119-25.

Breen J, Wicker T, Shatalina M, Frenkel Z, Bertin I, Philippe R, Spielmeyer W, Simková H, Safář J, Cattonaro F, Scalabrin S, Magni F, Vautrin S, Bergès H; International Wheat Genome Sequencing Consortium, Paux E, Fahima T, Doležel J, Korol A, Feuillet C, Keller B. A physical map of the short arm of wheat chromosome 1A. PLoS One. 2013;8(11):e80272.

Dabbert T, Okagaki RJ, Cho S, Heinen S, Boddu J, Muehlbauer GJ. The genetics of barley low-tillering mutants: low number of tillers-1 (lnt1). Theor Appl Genet. 2010 Aug;121(4):705-15.

Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009 Aug 1;25(15):1974-5.

Dibari B, Murat F, Chosson A, Gautier V, Poncet C, Lecomte P, Mercier I, Bergès H, Pont C, Blanco A, Salse J. Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. BMC Genomics. 2012 Jun 6;13:221.

Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011 Jan 31;6(1):e16526.

Gultyaev AP, van Batenburg FH, Pleij CW. The influence of a metastable structure in plasmid primer RNA on antisense RNA binding kinetics. Nucleic Acids Res. 1995 Sep 25;23(18):3718-25.

Hendriks PW, Kirkegaard JA, Lilley JM, Gregory PJ, Rebetzke GJ. A tillering inhibition gene influences root-shoot carbon partitioning and pattern of wateruse to improve wheat productivity in rainfed environments. J Exp Bot. 2016 Jan;67(1):327-40.

Ihsan MZ, El-Nakhlawy FS, Ismail SM, Fahad S, Daur I. Wheat Phenological Development and Growth Studies As Affected by Drought and Late Season High Temperature Stress under Arid Environment. Front Plant Sci. 2016 Jun 6;7:795.

Janssen BJ, Drummond RS, Snowden KC. Regulation of axillary shoot development. Curr Opin Plant Biol. 2014 Feb;17:28-35.

Jiao Y, Wang Y, Xue D, Wang J, Yan M, Liu G, Dong G, Zeng D, Lu Z, Zhu X, Qian Q, Li J. Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. Nat Genet. 2010 Jun;42(6):541-4.

Jinpeng Zhang, Jun Wu, Weihua Liu, Xiang Lu, Xinming Yang, Ainong Gao, Xiuquan Li, Yuqing Lu, Lihui Li Molecular Breeding. February 2013, Volume 31, Issue 2, pp 441-449 First online: 05 December 2012 Genetic mapping of a fertile tiller inhibition gene, ftin, in wheat.

Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 2000 Sep;16(9):418-20.

Kebrom TH, Spielmeyer W, Finnegan EJ. Grasses provide new insights into regulation of shoot branching. Trends Plant Sci. 2013 Jan;18(1):41-8.

Kronmiller BA, Wise RP. TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol. 2008 Jan;146(1):45-59.

Kruszka K, Pacak A, Swida-Barteczka A, Stefaniak AK, Kaja E, Sierocka I, Karlowski W, Jarmolowski A, Szweykowska-Kulinska Z. Developmentally regulated expression and complex processing of barley pri-microRNAs. BMC Genomics. 2013 Jan 16;14:34

Kuraparthy V, Sood S, Dhaliwal HS, Chhuneja P, Gill BS. Identification and mapping of a tiller inhibition gene (tin3) in wheat. Theor Appl Genet. 2007 Jan;114(2):285-94.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.

Li X, Qian Q, Fu Z, Wang Y, Xiong G, Zeng D, Wang X, Liu X, Teng S, Hiroshi F, Yuan M, Luo D, Han B, Li J. Control of tillering in rice. Nature. 2003 Apr 10;422(6932):618-21.

Lucas SJ, Budak H. Sorting the wheat from the chaff: identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. PLoS One. 2012;7(7):e40859.

Ma X, Xin Z, Wang Z, Yang Q, Guo S, Guo X, Cao L, Lin T. Identification and comparative analysis of differentially expressed miRNAs in leaves of two wheat (Triticum aestivum L.) genotypes during dehydration stress. BMC Plant Biol. 2015 Jan 27;15:21.

Mascher M, Jost M, Kuon JE, Himmelbach A, Aßfalg A, Beier S, Scholz U, Graner A, Stein N. Mapping-by-sequencing accelerates forward genetics in barley. Genome Biol. 2014 Jun 10;15(6):R78.

Mitchell JH, Rebetzke GJ, Chapman SC, Fukai S. Evaluation of reduced-tillering (tin) wheat lines in managed, terminal water deficit environments. J Exp Bot. 2013 Aug;64(11):3439-51.

Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P. Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet. 2004 Aug;109(4):800-5.

Pellegrineschi A, Noguera LM, Skovmand B, Brito RM, Velazquez L, Salgado MM, Hernandez R, Warburton M, Hoisington D. Identification of highly transformable wheat genotypes for mass production of fertile transgenic plants. Genome. 2002 Apr;45(2):421-30.

Pont C, Murat F, Guizard S, Flores R, Foucrier S, Bidet Y, Quraishi UM, Alaux M, Doležel J, Fahima T, Budak H, Keller B, Salvi S, Maccaferri M, Steinbach D, Feuillet C, Quesneville H, Salse J. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J. 2013 Dec;76(6):1030-44.

Prasad BD, Goel S, Krishna P. In silico identification of carboxylate clamp type tetratricopeptide repeat proteins in Arabidopsis and rice as putative co-chaperones of Hsp90/Hsp70. PLoS One. 2010 Sep 15;5(9):e12761.

Quraishi UM, Abrouk M, Murat F, Pont C, Foucrier S, Desmaizieres G, Confolent C, Rivière N, Charmet G, Paux E, Murigneux A, Guerreiro L, Lafarge S, Le Gouis J, Feuillet C, Salse J. Cross-genome map based dissection of a nitrogen use efficiency ortho-metaQTL in bread wheat unravels concerted cereal genome evolution. Plant J. 2011 Mar;65(5):745-56.

Quraishi UM, Murat F, Abrouk M, Pont C, Confolent C, Oury FX, Ward J, Boros D, Gebruers K, Delcour JA, Courtin CM, Bedo Z, Saulnier L, Guillon F, Balzergue S, Shewry PR, Feuillet C, Charmet G, Salse J. Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (Triticum aestivum L.). Funct Integr Genomics. 2011 Mar;11(1):71-83

Quraishi UM, Abrouk M, Bolot S, Pont C, Throude M, Guilhot N, Confolent C, Bortolini F, Praud S, Murigneux A, Charmet G, Salse J. Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. Funct Integr Genomics. 2009 Nov;9(4):473-84.

Raatz B, Eicker A, Schmitz G, Fuss E, Müller D, Rossmann S, Theres K. Specific expression of LATERAL SUPPRESSOR is controlled by an evolutionarily conserved 3' enhancer. Plant J. 2011 Nov;68(3):400-12.

Richards RA. (1988) A tiller inhibitor gene in wheat and its effect on plant-growth. Aust J Agric Res 39: 749–757

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell. 2008 Jan;20(1):11-24.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc Natl Acad Sci U S A. 2009 Sep 1;106(35):14908-13.

Satish Kumar, S. S. Singh, C. N. Mishra, Monika Saroha, Vikas Gupta, Pardeep Sharma, Vinod Tiwari, Indu Sharma. Assessment of Tiller Inhibition (tin) Gene Molecular Marker for its Application in Marker-Assisted Breeding in Wheat.

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, et al. Development and mapping of SSR markers for maize. Plant Mol Biol. 2002 Mar-Apr;48(5-6):463-81.

Spielmeyer W, Richards RA. Comparative mapping of wheat chromosome 1AS which contains the tiller inhibition gene (tin) with rice chromosome 5S. Theor Appl Genet. 2004 Oct;109(6):1303-10.

Suenaga K, Khairallah M, William HM, Hoisington DA. A new intervarietal linkage map and its application for quantitative trait locus analysis of "gigas" features in bread wheat. Genome. 2005 Feb;48(1):65-75.

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu JK. Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol. 2008 Feb 29;8:25.

Tavakol E, Okagaki R, Verderio G, Shariati J. V, Hussien A, Bilgic H, Scanlon MJ, Todt NR, Close TJ, Druka A, Waugh R, Steuernagel B, Ariyadasa R, Himmelbach A, Stein N, Muehlbauer GJ, Rossini L. The Barley Uniculme4 Gene Encodes a BLADE-ON-PETIOLE-Like Protein That Controls Tillering and Leaf Patterning. Plant Physiology. 2015/05/01 00:00; 168(1): 164-174

Unver T, Bakar M, Shearman RC, Budak H. Genome-wide profiling and analysis of Festuca arundinacea miRNAs and transcriptomes in response to foliar glyphosate application. Mol Genet Genomics. 2010 Apr;283(4):397-413.

Waldie T, McCulloch H, Leyser O. Strigolactones and the control of plant development: lessons from shoot branching. Plant J. 2014 Aug;79(4):607-22.

Wicker T, Yahiaoui N, Keller B. Contrasting rates of evolution in Pm3 loci from three wheat species and rice. Genetics. 2007 Oct;177(2):1207-16.

Wright M, Dawson J, Dunder E, Suttie J, Reed J, Kramer C, Chang Y, Novitzky R, Wang H, Artim-Moore L. 2001 Efficient biolistic transformation of maize (Zea mays L.) and wheat (Triticum aestivum L.) using the phosphomannose isomerase gene, pmi, as the selectable marker Plant Cell Reports 20 429–36.

You FM, Wanjugi H, Huo N, Lazo GR, Luo MC, Anderson OD, Dvorak J, Gu YQ. RJPrimers: unique transposable element insertion junction discovery and PCR primer design for marker development. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W313-20.

Zhang J, Zhang J, Liu W, Han H, Lu Y, Yang X, Li X, Li L. Introgression of Agropyron cristatum 6P chromosome segment into common wheat for enhanced thousand-grain weight and spike length. Theor Appl Genet. 2015 Sep;128(9):1827-37.

3. Discussion

3.1. Un microARN dicistronic contrôlerait l'inhibition du tallage chez le blé tendre.

Dans l'article précédent j'ai étudié la région chromosomique 1A, dite sensible au regard du processus de dominance des sous-génomes, porteuse du locus majeur TIN contrôlant l'inhibition du tallage, illustrant la diploïdisation des caractères post-polyploïdisation.

L'approche de génomique translationnelle en se basant sur les espèces apparentées, a permis d'affiner avec efficacité la cartographie génétique de la région du chromosome 1A impliquée dans le caractère du tallage. Sur la base de la modélisation de l'ordre des gènes de l'ancêtre des céréales AGK (le synténome), il a été possible de sélectionner des marqueurs moléculaires COS et cartographier génétiquement le locus grâce à une population SSD, issue du croisement entre les variétés Oligoculm et OC37. Ces mêmes marqueurs ont été utilisés pour cribler les banques BACs à disposition (Chinese Spring, Oligoculm et Renan), et initier la carte physique du locus. Après séquençage de ces fragments, le design de nouveaux marqueurs basés sur les éléments répétés (SSR et RJM) a permis de rechercher des recombinaisons au sein de populations de backcross entre Oligoculm et des variétés dites Elites (Apache, Caphorn et Koreli). La zone a été ainsi restreinte à 933 Mb de la séquence complète du chromosome 1A (séquence de référence NRgene, *cf.* page 9 Figure 6) porteuse de 18 gènes dont le re-séquençage a permis d'identifier une insertion de 109 pb présente chez 4 cultivars monotalle à notre disposition (2 israéliennes, 1 asiatique et 1 australienne). Cette insertion code pour un microARN dicistronic miRNA1436 spécifique des monocotylédones (Schreiber *et al.* 2011).

Les microRNA sont largement connus pour leur rôle dans la réponse à des stress abiotiques chez les plantes (Pandey *et al.* 2014 ; Sunkar *et al.* 2008, Kantar *et al.* 2010). Je propose dans cet article, qu'ils peuvent être des acteurs majeurs dans l'architecture du blé tendre. Des études ont déjà confirmées que cette famille de gènes peut contribuer aux phénomènes de néo- et sousfonctionnalisation de leurs cibles. Li *et al.* 2014 ont montré que les miRNAs généraient une non-additivité à hauteur de 24 % chez de jeunes épis de blés synthétiques comparés à la valeur de leur géniteurs tétraploïdes et diploïde ; avec une très forte répression (75 %) des miRNA hérités de *A. tauschii* (génome D) en réponse aux stress abiotiques (sécheresse, sels & froid). Ces microARN en structure tige/boucle peuvent être facilement source de plasticité. En effet, une simple mutation peut dénaturer la fonction de régulation du microARN. De plus, différents précurseurs peuvent générer le même miRNA mature permettant une expression différentielle selon les conditions spatio-temporelles. Provenant parfois d'éléments transposables tels que les MITEs, ils sont abondamment répartis sur le génome (Kruszka *et al.* 2013 ; Lucas *et al.* 2012). Ainsi, ils sont présents dans les régions non codantes où la diversité nucléique connait un turn-over important (Roffler *et al.* 2015), notamment *via* les transposons de type Mariner (l'activité des transposons ADN étant une force majeure de l'évolution générant de la diversité génétique).

Chez Oligoculm, l'allèle TIN semble être constitué d'un microARN dicistronique par comparaison avec les autres variétés cultivées où ce microARN est exclusivement monocistronique. Ce type de microARN polycistronique a été identifié chez les plantes et mammifères avec des clusters parfaitement homologues (cas de TIN) ou non-homologues (Baldrich *et al.* 2016). Chez les plantes, il semble que 20 %

des miRNA sont sous forme de clusters homologues (Merchan *et al.* 2009). Chez le riz, OSA-miR7695 est un miRNA polycistronique qui a évolué récemment suite à la sélection naturelle et/ou la domestication. Sa surexpression confère une résistance aux agents pathogènes (Campo *et al.* 2013). L'avantage de ce mécanisme est qu'il peut permettre de transcrire simultanément plusieurs miRNA (de la même famille ou non) *via* un seul précurseur qui sera par la suite clivé, débouchant ainsi à une co-régulation post-transcriptionnelle de cibles potentiellement différentes.

3.2. Le locus TIN, un exemple de diploïdisation de caractère

Ces travaux constituent une illustration de la diploïdisation phénotypique en réponse à la diploïdisation fonctionnelle et structurale documentée dans les chapitres précédents. Ce locus, positionné sur le chromosome 1A, fait partie du groupe chromosomique 1 (chromosomes 1A, 1B, 1D) dit sensible, suite à la duplication ancestrale A1/A5, il y a 90 MYA (*cf*. Figure 53). La région paralogue du groupe chromosomique 3 (chromosomes 3A, 3B, 3D) est, elle, dite dominante. Le miRNA1436 (ou de manière générale l'insertion de 109 bp) est spécifique du chromosome 1A. Le gène flanquant nous renseigne sur un possible scénario évolutif de ce locus et de la diploïdisation de ce caractère (Figure 54). Alors que de nombreux gènes homéologues peuvent être identifiés entre les chromosomes 1A, 1B, 1D et 3A, 3B, 3D, seulement 2 gènes sont conservés au locus TIN entre ces deux groupes chromosomiques ancestraux dupliquées il y a 90 MYA (Figure 54, ligne rouge et ligne verte). Parmi ces 2 gènes, on retrouve le gène Ankyrin conservé chez *Brachypodium* (Bradi2g39890), en amont de l'allèle TIN portant le miRNA1436 (partie basse de la Figure 54). A l'échelle de la micro-synténie (au niveau génique), on peut s'apercevoir que seulement une petite partie du gène est conservée entre la région 3B et 1A, 1B, 1D. Une très bonne conservation de la structure génique est retrouvée avec *Brachypodium* (Bd) mais faible avec *Arabidopsis thaliana* (Ar). Ce Locus est donc potentiellement ancestral car il est conservé chez *Brachypodium* notamment, puis il a été perdu chez le blé tendre entre les deux groupes chromosomiques dupliqués (sur les sous-génomes 3A et 3D). Au-delà de cette diploïdisation des paralogues ancestraux, une diploïdisation est également identifiable entre les homéologues du groupe chromosomique 1 avec des séquences UTR distinctes ; la région 1A étant la seule pourvue du miRNA1436 en région 5' UTR, ce miRNA1436 étant alors dicistronique (insertion de 109 bp) pour les variétés monotalles. Cet exemple illustre parfaitement les différents niveaux de plasticité chez le blé tendre entre régions dupliquées ancestralement (ici entre régions ancestrales A1 et A5 qui correspondent aux groupes chromosomiques 3 et 1 du blé tendre), entre les régions récemment dupliquées (ici entre les sous-génomes A, B et D) et enfin entre variétés (ici entre génotypes TIN+ et TIN-).

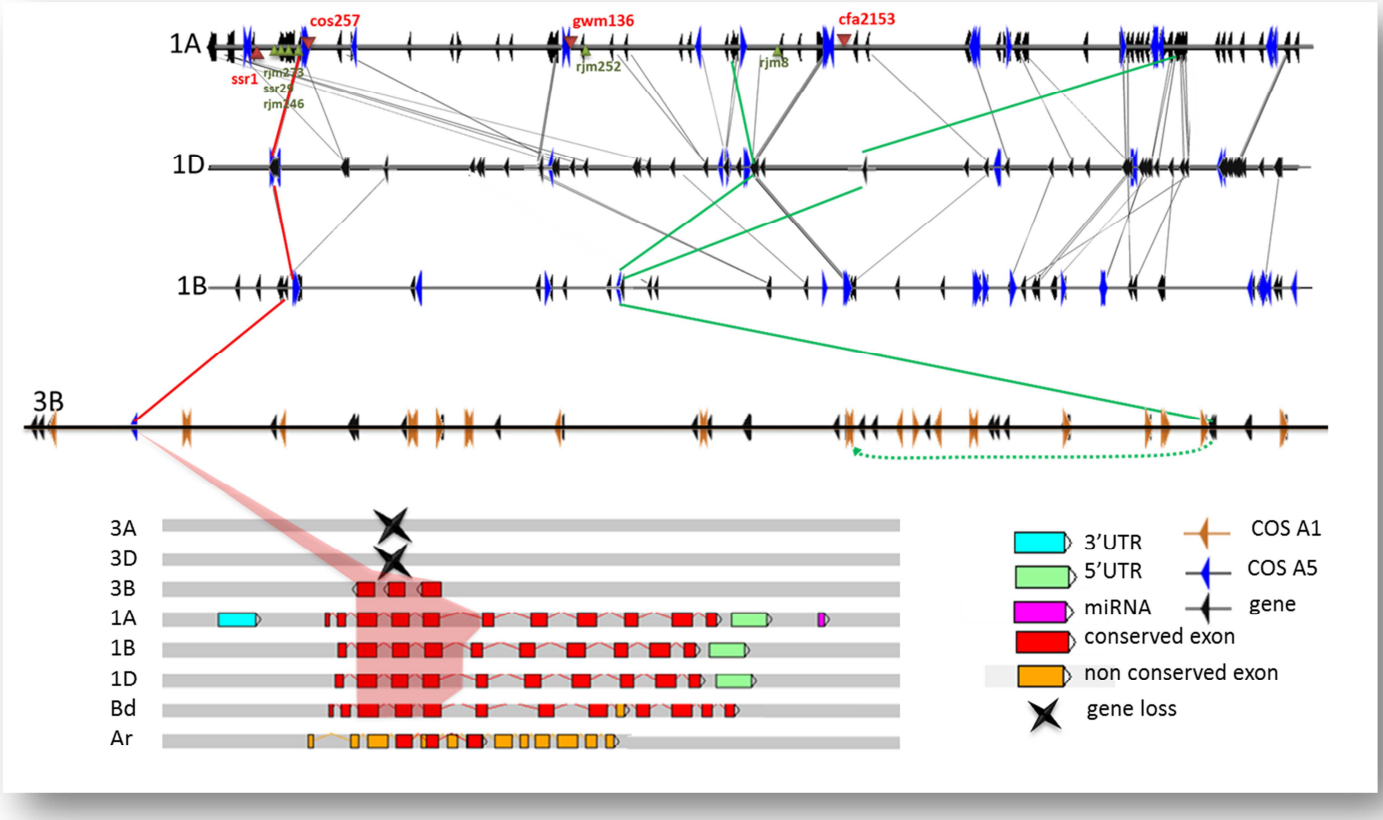


Figure 54. Régions paralogues/homéologues du locus TIN chez le blé tendre.
En haut de la figure les 3 régions homéologues sont représentées 1A, 1D puis 1B (3,5 Mbases chacune). Les marqueurs moléculaires sont matérialisés par des triangles sur la séquence 1A. Les gènes annotés en bleu sont conservés chez les céréales apparentées (en provenance du chromosome ancestral A5) et les gènes annotés en noir sont spécifiques du blé tendre. Les traits matérialisent les relations d'homologie entre les gènes. La région paralogue 3B (5Mbases) est représentée dessous avec les gènes annotés en orange conservés chez les céréales apparentées et provenant du chromosome ancestral A1. Les gènes annotés en noir sont spécifiques du blé tendre. Seuls 2 gènes sont conservés entre ces 2 régions ancestrales. Au bas, la micro-synténie de la région conservée ancestralement correspondant au marqueur COS257, est représentée avec les exons conservés (en rouge) chez *Arabidopsis thaliana* (Ar), *Brachypodium* (Bd) et sur les régions du blé tendre 3B et 1A, 1B, 1D. La croix matérialise la délétion du gène.

3.3. Impact de l'allèle TIN sur la capacité adaptative de la plante

Au cours de l'évolution, les allèles favorables à un meilleur succès reproducteur dans un environnement donné sont conservés *via* la sélection naturelle. Ainsi, le patrimoine génétique d'une plante qui s'adapte aux contraintes environnementales (et par essence survit) sera transmis à la descendance au sein de la population. A ce titre Mitchell *et al.* 2013 décrivent le rôle potentiel de l'allèle TIN dans l'aptitude à la résistance au stress hydrique. Alors que la sécheresse diminue le tallage, augmente le taux d'avortement des épillets et favorise l'apparition de petits grains, les lignées portant le locus TIN (TIN+) produisent des épis plus résistants à la sécheresse en maintenant le remplissage du grain (+11 %). Cette résistance est associée à une augmentation de la biomasse à la floraison et à la disponibilité de sucres dans la tige pouvant par la suite être mobilisés lors du remplissage du grain. Les auteurs suggèrent, que même à

densité moyenne, l'allèle contribue à une augmentation du rendement (Mitchell *et al.* 2013, Duggan *et al.* 2006). Moeller *et al.* 2014 décrivent également un retard de sénescence du couvert végétal, lié à la disponibilité accrue de la lumière, de l'eau et de l'azote par tige. Les racines contribuent également à cette architecture spécifique avec un ratio racine/tige augmenté, dès le début du développement de la plante portant l'allèle TIN, retardant l'utilisation de l'eau contenue dans le sol. Ainsi, la disponibilité de la plante en eau est prolongée en raison d'une meilleure conductance stomacale (échange gazeux) et de la transpiration, permettant le maintien 'vert' du couvert végétal et ainsi *in fine* la possible résistance au stress hydrique (Chimungu *et al.* 2014).

4. Perspectives

Validation fonctionnelle - Nous avons entrepris la validation fonctionnelle du miRNA1436 dicistronique potentiellement responsable de l'inhibition du tallage chez Oligoculm. Pour cela nous avons choisi la méthodologie CRISPR-Cas9 (Sugano *et al.* 2014) pour inactiver l'allèle TIN+ (mutations spécifiques ou délétion de tout ou partie de l'insertion de 109 bp) d'Oligoculm et redonner ainsi à Oligoculm la capacité à taller. Cette biotechnologie utilisent des protéines capables de couper l'ADN, comme des ciseaux à un endroit précis, pour y induire une/des mutation/s, ou générer une délétion. Des ARN guides spécifiques de miRNA1436 permettent de guider la protéine Cas9 jusqu'au locus à couper. La construction génétique a été introduite dans des cellules végétales de la variété Oligoculm par biolistique (bombardement de billes d'or sur des embryons immatures, 12–14 jours après floraison) comme décrit par Pellegrineschi *et al.* 2002, par la plateforme de validation fonctionnelle de l'UMR INRA-UBP GDEC (www6.clermont.inra.fr/umr1095/Equipes). La construction utilisée est spécifique de l'insertion d'Oligoculm (CRISPR-Cas plasmid L1509 Cas9 rice + puc57 + gRNAtin1 + gRNAtin2 + pmi) avec pour guides gRNAtin1: GGATAGGGCTCAATTTGCGTCGG et gRNAtin2: ACGTTTACATTATGGGACGGAGG. La transformation, la régénération, la sélection (selon Wright *et al.* 2001) et l'analyse moléculaire des plantes transformées a permis de caractériser 27 évènements indépendants retenus en T0 pour lesquels les grains ont été récoltés en 2015. En 2016, pour chaque famille, 16 grains ont été semés et les plantes séquencées avec les primers diagnostics F24R23. Aucune mutation n'a pu être détectée sur les générations T1 criblées.

Sur la base de cet échec, nous avons par la suite entrepris la validation fonctionnelle du locus TIN par surexpression de l'allèle Oligoculm (l'insertion de 109 bp) dans le cultivar Courtot, à fort tallage, pour tenter de lui faire acquérir la capacité à inhiber son tallage. La construction plasmidique utilisée (plasmid pDESTR4-R3 ubi promotor + 109pbtin + nosubipminos) permet de faire exprimer fortement l'allèle Oligoculm à l'aide d'un promoteur ubiquitaire. En 2016, 15 familles sont étudiées en T1 (notation du nombre de talles) à raison de 16 plantes par famille après vérification de la présence de l'insert par PCR. Ce travail est actuellement en cours et constitue la piste prioritaire pour la validation fonctionnelle de l'allèle TIN+, prérequis avant la soumission de l'article précédent pour publication.

Matériel TIN+ pour le pre-breeding - L'introgression de l'allèle TIN dans des schémas de sélection peut être intéressante pour le maintien du rendement en condition de faibles intrants *via* une meilleure

tolérance en carence azotée voire en condition de stress hydrique. Ce caractère est particulièrement recherché en agriculture biologique et en agriculture raisonnée qui vise un itinéraire technique économe. De plus, un semis plus tardif associé à une densité réduite, s'accompagne d'un faible besoin d'azote au moment du tallage. Tout ceci fait que l'allèle Oligoculm est intéressant à étudier en breeding. Pour accélérer ce processus visant à utiliser l'allèle TIN dans les programmes de sélection, nous avons délibérément introduit le locus TIN par backcross dans des fonds génétiques dits 'élite' ; Apache, Koreli et Caphorn (discutés dans l'article). Apache est couramment utilisé ; inscrit au catalogue depuis 1997, il a servi de témoins au sein des essais CTPS (comité technique permanent de la sélection) entre 2006 et 2009. Il représente 3,9 % des semences multipliées en 2014, surtout du fait de sa résistance à la fusariose et de sa bonne teneur en protéines. Caphorn lui, est plus économique, car il a un meilleur rendement qu'Apache, mais une teneur en protéines moyenne. Koreli est plus tardif, mais possède un fort poids spécifique du grain et représente 1,1 % des semences multipliées en 2013 (inscription en 2005), de plus il est assez résistant aux maladies en conditions non-traitées.

Dans le cadre d'un projet financé par le fond de soutien à l'obtention végétale (FSOV), les lignées d'introgression décrites dans ce chapitre, sont en cours d'évaluation en réponse à différents apports azotés et à différentes densités de semis. Ainsi en 2015 et 2016, les lignées ont été multipliées pour satisfaire les besoins en grains de l'étude d'évaluation en plein champ. Lors de la campagne 2016, j'ai réalisé des notations de phénotypage sur ces lignées afin de les comparer aux valeurs parentales pour le nombre de talles par plante (*cf*. figure 55, histogrammes à gauche). Les lignées d'introgression que nous avons construites permettent de disposer, dans un fonds génétique élite, d'une grande gamme de tallage (du monotalle Caphorn à gauche, à 8 talles pour Koreli à droite, Figure 55) permettant d'étudier la réponse de ce matériel à différents itinéraires culturaux ainsi que différentes contraintes biotiques et abiotiques.

En effet, la résistance à la sécheresse de ces lignées pourrait également être étudiée en situation agronomique (plein champ) mais nécessite un dispositif conséquent. Il serait envisageable de tester leurs performances avec la plateforme Pheno3C (site de Crouël), intégrée à l'unité PHACC (Unité Expérimentale PHénotypage Au Champ des Céréales), qui permet d'appliquer des méthodes de phénotypage haut débit au champ, en condition de stress hydrique.

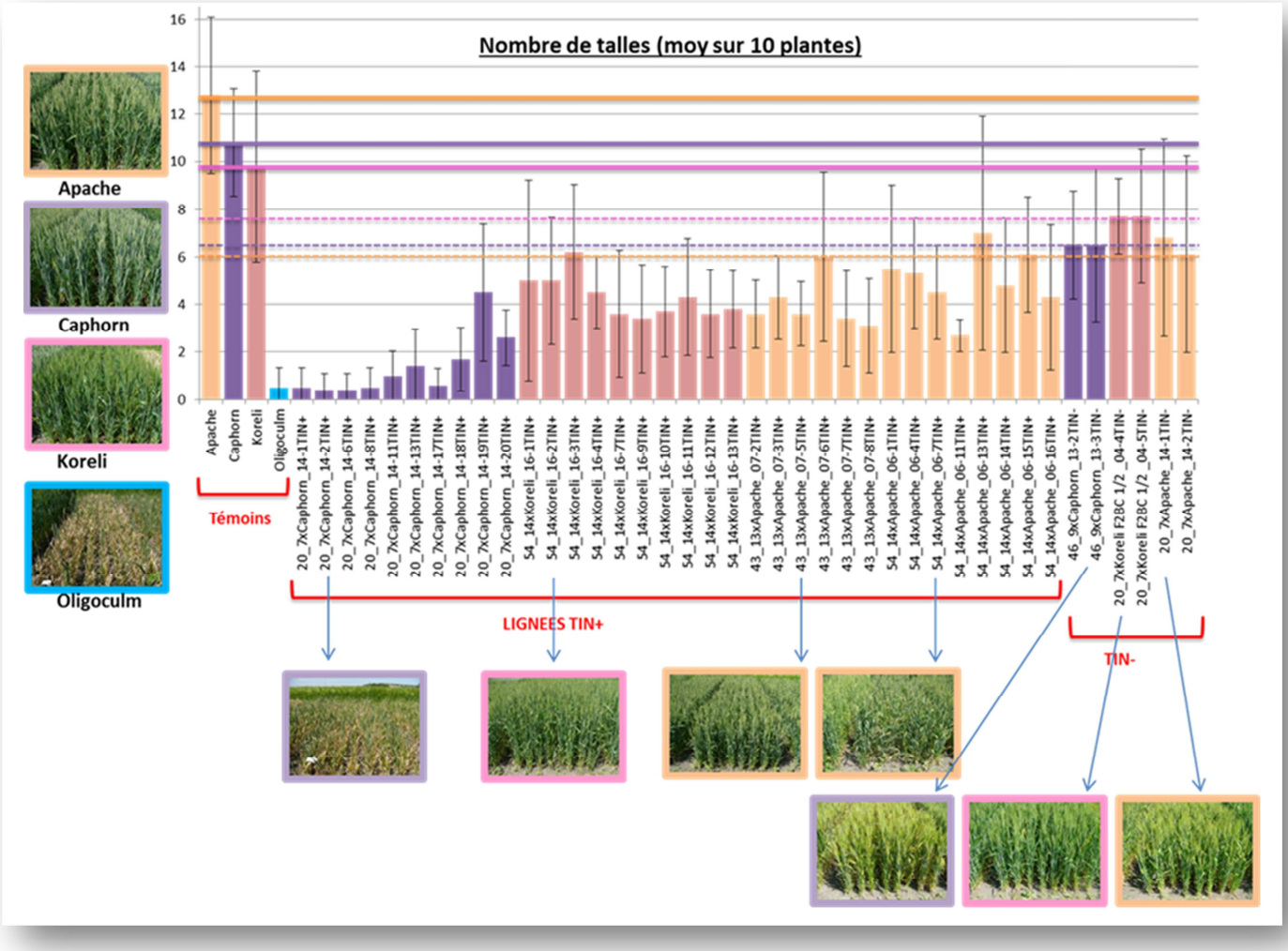


Figure 55. Distribution du tallage de lignées d'introgression portant l'allèle TIN

La figure illustre la distribution du tallage des lignées backcrossées selon les fonds génétiques ; Apache en orange, Caphorn en violet et Koreli en rose. Le tallage d'Oligocolum est matérialisé en bleu. Le nombre de talle par plante a été comptabilisé sur 10 individus de chaque lignée. Les photos illustrent la densité de la parcelle.

5. Conclusion

La recherche fondamentale présentée en partie 1 et 2 peut avoir une application directe, les travaux sur le tallage en sont un exemple concret. En effet, le tallage, ou capacité de la plante à produire des épis, est une composante forte du rendement permettant de maintenir un rendement élevé même sous des contraintes environnementales non favorables. Ainsi, toutes les connaissances acquises sur la plasticité du génome du blé tendre associées aux informations connues chez les espèces modèles de céréales apparentées (riz, maïs et sorgho) ont permis d'étudier la diploïdisation du caractère de tallage porté par le chromosome 1A chez le blé tendre. J'ai étudié la région chromosomique dite sensible 1A, porteuse du locus impliqué dans le tallage. L'approche intégrative de génomique translationnelle, robuste et rapide, a permis de caractériser le gène d'intérêt et de l'introgresser dans des variétés commerciales qui seront évaluées au champ et pourront être introduites dans le programmes de pre-breeding.

Ainsi, ce dernier volet montre, qu'à l'image de la diploïdisation structurale et fonctionnelle, la diploïdisation des caractères est retrouvée à l'échelle du blé tendre issu d'une polyploïdisation il y a 500 000 et 10 000 ans. Cette diploïdisation des caractères ou spécialisation phénotypique des sous-génomes post-polyploïdie a possiblement permis au blé d'être cultivé dans une large aire géographique. Ainsi, il occupe avec succès divers habitats, en affichant une multitude de facettes morphologiques, biochimiques et moléculaires que les espèces progénitrices diploïdes ne possèdent pas.

Ces derniers travaux démontrent également la pertinence de la biologie translationnelle pour travailler des caractères agronomiques à partir des espèces modèles, même s'ils sont diploïdisés. Que ce soit pour des caractères conservés ou spécifiques d'espèces, les outils, ressources et méthodologies développés lors de cette approche permettent de cibler avec efficacité la région d'intérêt (*via* la cartographie génétique et physique), d'affiner l'annotation des gènes de l'espèce d'intérêt et mieux sélectionner les gènes/loci candidats. Cette méthode a déjà fait ses preuves pour la caractérisation du locus NUE (Nitrogen Use Efficiency ; Quraishi *et al.* 2011a), la teneur en fibre du grain (Quraishi *et al.* 2011b), la teneur en caroténoïde du grain (PSY, Dibari *et al.* 2012), l'allèle du blé miracle FRIZZY PANICLE (FZP ; Dobrovolskaya *et al.* 2015) auxquels j'ai été associée ces dernières années et ici le tallage, qui fait partie de mes travaux de thèse.

Nous avons montré que la diploïdisation structurale et expressionnelle des sous-génomes, après les événements de polyploïdie, est suivie par une diploïdisation phénotypique. Les sous-génomes deviennent alors, le lieu d'une spécialisation de caractères non-présents dans les progéniteurs diploïdes, permettant ainsi, potentiellement, une meilleure adaptation à l'environnement et une tolérance accrue aux contraintes.

Conclusions & perspectives

CONCLUSIONS

1. Contribution des travaux de thèse à l'étude de l'impact de la polyploïdie

Les travaux récents effectués au sein de l'équipe, ont contribué à l'étude de l'histoire évolutive des génomes de céréales à partir d'un ancêtre commun il y a 90 MYA. Il a ainsi été établi que la polyploïdie est un évènement récurrent de l'évolution des plantes. Il apparaît acquis que la polyploïdie constitue un 'choc génomique' qui permet une 'reprogrammation' des génomes post-polyploïdie. La polyploïdie est suivie d'une diploïdisation structurale (perte des gènes) et fonctionnelle (modification de l'expression des gènes) conduisant au phénomène de dominance des sous-génomes (dérivant des compartiments dominants et sensibles). Ces phénomènes ont été caractérisés sur des espèces modernes ayant subie des polyploïdisations anciennes, de telle sorte que ces espèces sont aujourd'hui diploïdes.

La dominance des sous-génomes existe-t-elle chez les polyploïdes modernes ? Telle était la question initiale de ces travaux de thèse. Le blé hexaploïde (AABBDD) constitue, dans ce contexte scientifique, un excellent modèle permettant l'étude de l'impact de la polyploïdie sur la structure et la fonction des gènes, car celui-ci a subi des évènements anciens et récents de polyploïdie il y a 90 MYA et jusqu'à 10 000 ans (paléo et néo-polyploïdie). Même si la communauté scientifique ne dispose pas encore des 21 pseudomolécules exhaustives du génome du blé tendre, de nombreuses ressources génomiques sont disponibles permettant l'étude de l'impact de la polyploïdie sur l'organisation et la régulation du génome du blé tendre moderne.
La dominance des génomes est une réponse à la polyploïdie qui induit un retour à l'état diploïde compartimenté. A travers mes travaux de thèse (Chapitre I), cette diploïdisation a pu être caractérisée et quantifiée chez le blé hexaploïde. Une diploïdisation structurale engagée pour 38 % des gènes (avec la perte d'au moins un homéologue) et avec une sensibilité différentielle des sous-génomes B>A>D (*i.e.* conservation contrastée des gènes ancestraux respectivement de 62 %<66 %<69 %). La paléo-dominance (en réponse à la duplication ancestrale) se surimpose à la polyploïdie récente (hexaploïdie) pour donner naissance aux compartiments supra-dominants (D-D ; fraction dominante du génome D avec forte rétention de gènes ancestraux) et supra-sensibles (S-B ; fraction sensible du génome B avec une faible rétention de gènes ancestraux). Les travaux de cette thèse permettent ainsi de généraliser le processus de dominance des sous-génomes post-polyploïdie en réponse à l'hexaploïdie récente avec une plasticité accélérée (ou sensibilité) du sous-génome B (Figure 56).

Mes travaux (chapitre II) ont également permis de démontrer une différenciation massive de l'expression des gènes dupliqués chez le blé tendre moderne. Au regard de la polyploïdie et au sein des 3 génomes, moins de 10 % des gènes homéologues (triplets A, B et D) présentent le même profil d'expression au cours du développement du grain et 33 % de ces triplets ont 3 copies complètement divergentes, après moins de 500 000 ans. De même, si l'on s'intéresse à la duplication ancestrale datant de 90 MYA, ayant produit 6 copies potentielles chez le blé tendre, on parle alors de gènes ohnologues, uniquement 20% d'entre eux ont conservé l'expression ancestrale (c'est-à-dire 2 copies sur 6 issues de la WGD avec la même expression).

Enfin, mes travaux suggèrent, au travers de la dissection des bases génétiques du tallage, que la diploïdisation structurale et expressionnelle des sous-génomes après les événements de polyploïdie est suivie par une diploïdisation phénotypique. Les sous-génomes deviennent alors, le lieu d'une

spécialisation de caractères non-présents chez les progéniteurs diploïdes, permettant potentiellement une adaptation aux contraintes environnementales.

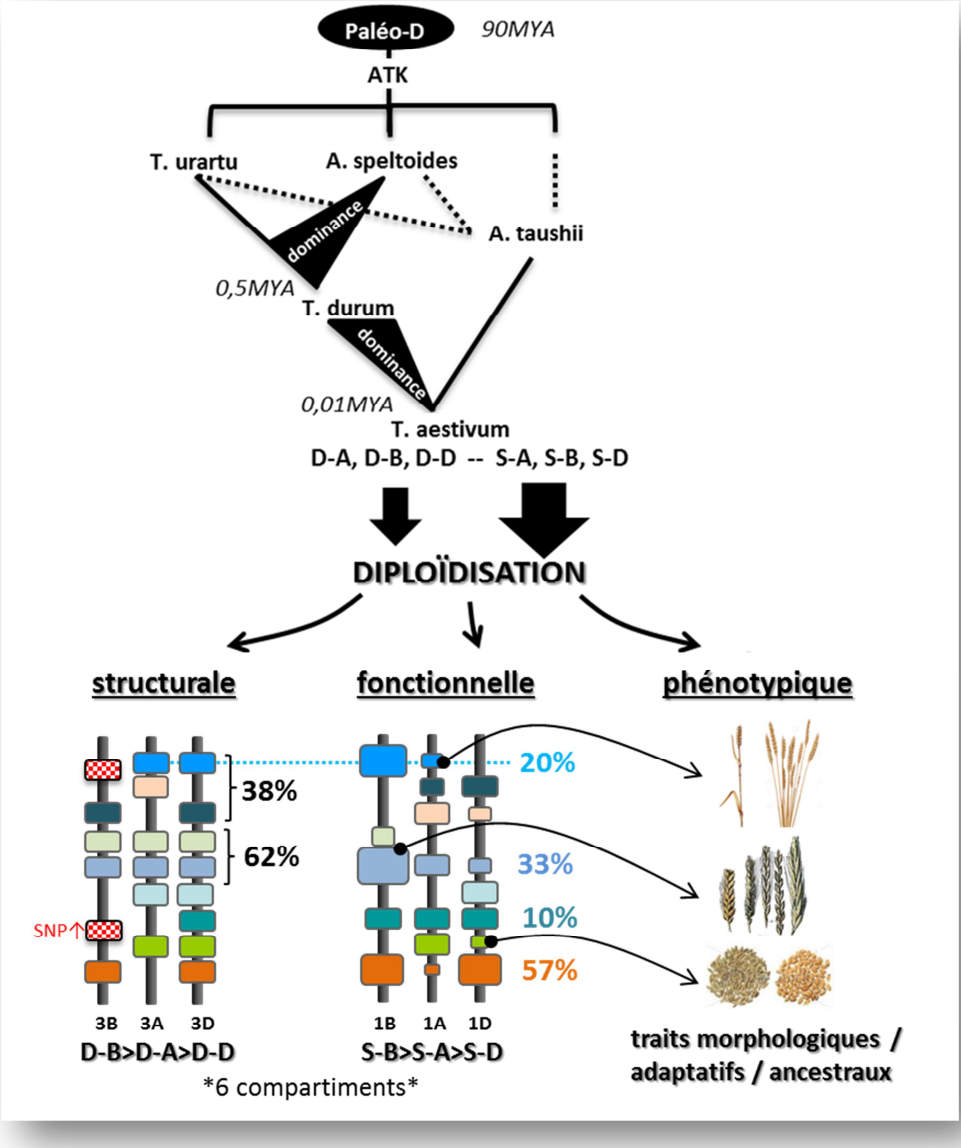


Figure 56. Diploïdisation structurale, fonctionnelle et *phénotypique chez le blé*

L'histoire évolutive du blé est schématisée à partir de l'ancêtre des céréales ATK ayant déjà subi une duplication (90MYA) avant de retourner à l'état diploïde sous l'effet de la paléo-dominance des blocs D et S. Les 3 génomes diploïdes se sont hybridés récemment (<0,5 MYA) pour donner le blé tendre, soumis lui, à une néo-dominance entre les sous-génomes A, B et D. Les duplications combinées (paleo-tétraploïdie et neo-hexaploïdie) ont fait naître chez ce blé moderne 6 compartiments génomiques dupliqués (D-A; D-B; D-D; S-A; S-B; S-D). Les 3 types de diploïdisation (structurale, expressionnelle et phénotypique) sont illustrés en bas, en prenant à titre d'exemple la diploïdisation structurale sur le groupe chromosomique dominant 3 et la diploïdisation fonctionnelle sur le groupe chromosomique sensible 1. A droite est représentée la spécificité des caractères agronomiques apportée par un seul sous-génome (diploïdisation phénotypique). Au niveau structural, 38% des gènes sont affectés par la perte de copies redondantes. Le sous génome B caractérisé par une plasticité nucléique accélérée est matérialisé avec l'apparition de gènes riches en SNPs (en rouge). Au niveau fonctionnel, moins de 10% des copies homéologues ont une redondance d'expression, et 33% ont une divergence totale (des trois copies). 20 % des ohnologues ont conservé au moins un profil ancestral (en bleu) sur les deux blocs D et S. Au niveau phénotypique, de nombreux traits morphologiques, adaptatifs et ancestraux sont apportés par des allèles homéologues spécifiques.

2. Questions soulevées par les travaux de thèse sur le mécanisme de dominance des sous-génomes post-polyploïdie

La littérature ouvre de nombreuses réflexions concernant les mécanismes impliqués dans la dominance des sous-génomes observée ; avec notamment l'hypothèse d'une origine épigénétique qui est souvent émise (Doyle *et al.* 2008, Tittel-Elmer *et al.* 2010, Wang *et al.* 2016). Dans le paragraphe suivant, je suggère des pistes pour rechercher les mécanismes impliqués dans la dominance des sous-génomes.

2.1. Relation entre épigénétique et dominance des sous-génomes

La forte présence d'hétérochromatine préférentiellement au niveau du sous-génome sensible B pourrait refléter des différences de marques épigénétiques au niveau des histones. L'étude de Wang *et al.* 2016 indique une implication du gène lysine-specific histone demethylase 1 (LSD1) dans l'effet trans-générationnel déclenché suite à un stress thermique appliqué à une lignée parentale de blé. Par ailleurs, comme déjà discuté dans le chapitre II (Perspectives), la méthylation de l'ADN peut être en lien avec le phénomène de dominance des sous-génomes. En effet, l'allopolyploïdie peut induire des changements de méthylation au niveau des transposons (chez les spartines ; Parisod *et al.* 2010) et au niveau des gènes (chez le blé ; Chagué et al. 2010). Le taux de transitions (SNP de type [C => T]) peut aussi refléter des traces ancestrales de la méthylation de l'ADN. Les dinucléotides CpG méthylés peuvent dans ce contexte conduire à la mutation de la cytosine en thymine (Xia *et al.* 2012).

Sur cette base, il semblerait pertinent d'approfondir la distribution de l'ADN méthylé, le taux de transition C/T ainsi que les marques des histones, et ce, au regard de la compartimentation (les 6 compartiments issus de la paléo-tétraploïdie et de la néo-hexaploïdie) du génome du blé. Chez les plantes les zones sujettes au silencing par la méthylation de l'ADN sont notamment les zones répétées à fortes teneur en éléments transposables. L'étude du contexte en TE des gènes peut être intéressante lors de l'analyse des méthylations de l'ADN (et siRNA) car l'activation d'une famille de rétrotransposons (copia Onsen) a été observée suite à un stress thermique (Tittel-Elmer *et al.* 2010) chez A*rabidopsis thaliana*. Chez le blé, l'étude du BS-seq (bisulfite sequencing) au niveau du génome entier est encore difficilement envisageable du fait de la taille de son génome et de sa teneur en TE. Pour pallier cela, la méthylation d'ADN peut être investie à travers la capture de régions génomiques (*via* l'hybridation de sondes en milieu liquide) suivie du traitement bisulfite des bases méthylées avant séquençage (Lee *et al*. 2011). Quant au statut épigénétique, il pourrait être exploré chez le blé, *via* des marques présentes dans l'euchromatine avec par exemple les marques H3K9ac associées à la transcription des gènes et H3k27m3 à leur répression. Le rôle de ces modifications épigénétiques en lien avec la dominance des sous-génomes post-polyploïdie sera étudié en détail au laboratoire à partir des blés synthétiques décrits dans la Chapitre II (page 65).

2.2. Impact des éléments transposables

Perte de gènes - Certains mécanismes de pertes de gènes ont été étudiés chez le maïs, montrant que le mécanisme d'élimination des transposons est impliqué dans le fractionnement du génome (Woodhouse *et al.* 2010, Freeling *et al.* 2012), voire moteur de celui-ci. Les auteurs ont montrés que dans un nombre significatif de cas, l'un des 2 gènes dupliqués comporte une délétion d'exon. L'analyse des séquences répétées flanquantes chez les espèces orthologues démontre que le gène redondant est supprimé par recombinaison intrachromosomique locale (Woodhouse *et al.* 2010, *cf.* Figure 57). Ainsi, ce mécanisme induirait sur les chromosomes sensibles la perte ou pseudogénisation des gènes.

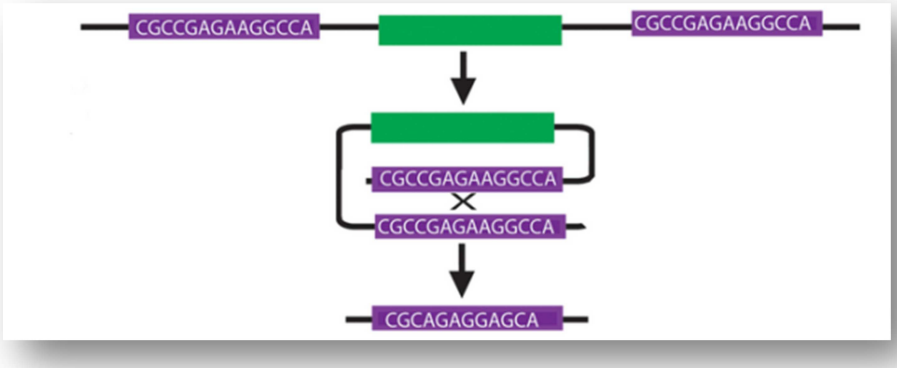


Figure 57. Mécanisme de recombinaison intra-chromosomique à l'origine de la perte de gènes.
Représentation schématique du mécanisme de recombinaison intra-chromosomique basé sur un appariement de régions répétées. Un gène est matérialisé en vert, encadré par 2 régions identiques flanquantes (en violet). Les éléments répétés se circularisent ; lors de la recombinaison une seule répétition flanquante subsiste, supprimant en même temps tout ou partie du gène. *Source : Woodhouse et al. 2010*

Les pseudogènes peuvent donner lieu à de petits ARN (24nt–siRNA) réprimant ainsi leur transcription (*cis*-acting) mais aussi induire le silencing en interagissant avec le transcrit complémentaire du gène parental fonctionnel (*trans*-acting ; Guo *et al.* 2009). Il a été vu chez le riz, que les siRNAs antisens peuvent être produits à des stades développementaux et conditions de cultures spécifiques, ce qui suggère un rôle majeur dans la plasticité phénotypique (Guo *et al.* 2009).
Chez le riz, la classe 2 des éléments transposables est minoritairement retrouvée sur le sous-génome sensible (*cf.* Chapitre I ; article Pont et al 2013). Cette classe 2, ou transposons, est connue sur le mode du « couper-coller », c'est-à-dire que leur transposition est couplée à l'excision du site d'origine. Cette excision peut capturer un gène (tout, ou en partie), induisant sa transposition, sa délétion ou sa pseudogénisation.

Expression des gènes – Sachant que l'expression des gènes est corrélée négativement avec la densité de TE méthylés (Hollister *et al.* 2009) et que les transposons sont retrouvés minoritairement sur les génomes sensibles, existe-t-il un lien entre dominance, TE méthylés et expression des gènes ? L'expression non-additive des gènes dans le contexte de polyploïdes synthétiques est régulée post-

transcriptionnellement *via* les petits ARNs ; microRNAs, small interfering RNAs (siRNAs), et *trans*-acting siRNAs (tasiRNAs, Baulcombe *et al.* 2004). Les microRNAs sont produits à partir d'un locus indépendant de leur cible et servent de répresseurs de l'expression des gènes, par la dégradation des ARN. Ainsi, la combinaison (ou interaction) de microRNAs et de leurs cibles provenant de différents géniteurs (dans un hybride ou dans un polyploïde), peut reprogrammer le génome. Les petits ARNs de type repeat-associated siRNAs (rasiRNAs) sont peu conservés et majoritairement issus des TE des régions hétérochromatiniennes (Lippman *et al.*2004). D'après Chen *et al.* 2010, chez *A. thaliana* la population rasiRNAs est relativement faible en F1 (issu du croisement *Arabidopsis thaliana Columbia* et *A. thaliana* C24), et de nombreux rasiRNAs absents en F1 sont restaurés plus tard dans l'allotétraploïde naturel (issu de l'hybridation entre *A. thaliana* Ler et *A. arenosa)*. Il semble falloir plusieurs générations pour établir des profils d'expression stables de siRNA régulant des protéines. Par contre concernant les miRNA, la proportion est plus élevée en F1, indiquant des changements rapides et dynamiques dès la polyploïdisation. Au fil du temps, la stabilité du génome est rétablie grâce à la régénération des rasiRNAs dans les allotétraploïdes génétiquement stables.

Certains auteurs (Song *et al.* 2010, Chen *et al.* 2010, Mosher *et al.* 2009) font le lien entre petits ARN et empreintes parentales lors de la fécondation. Selon Chen *et al.* 2010, une hypothèse probable est que les siRNA hérités de la mère, *via* le cytoplasme, répriment les transposons qui se sont réactivés durant la gamétogenèse. Ainsi, le « choc génomique » des hybrides interspécifiques peut provoquer une instabilité du génome *via* les TE. Sur cette base il serait pertinent de mener des études complémentaires sur la caractérisation des TE (ainsi que leur niveau de méthylation) chez le blé tendre, mais aussi en petits ARNs, en réponse à la polyploïdisation. Le rôle potentiel des TE en lien avec la dominance des sous-génomes post-polyploïdie sera étudié en détails au laboratoire à partir des blés synthétiques. Pour cela, le « mobilome » sera étudié, c'est-à-dire la fraction en TE remobilisée et active lors des premières générations post-polyploïdie (Collaboration avec Marie Mirouze, Laboratoire LGDP de Perpignan).

2.3. Phénomène d'hétérosis

Dans la nature, l'allopolyploïde présentant une forte « vigueur » remplace parfois les lignées parentales et colonise totalement son environnement, à l'image de *Spartina townsendii* qui s'est propagé sur les côtes du Royaume-Uni et en France. Une question légitime peut se poser, à savoir si l'effet de la polyploïdie (et des mécanismes de la dominance des sous-génomes associée) sont les mêmes que ceux impliqués dans les phénomènes d'hétérosis chez les hybrides. Le phénomène d'hétérosis est décrit comme une meilleure vigueur d'un hybride par rapport à ses parents (supériorité en termes de biomasse, stature, développement) et peut légitimement rappeler l'effet phénotypique obtenu chez les polyploïdes (*cf.* Figure 58).

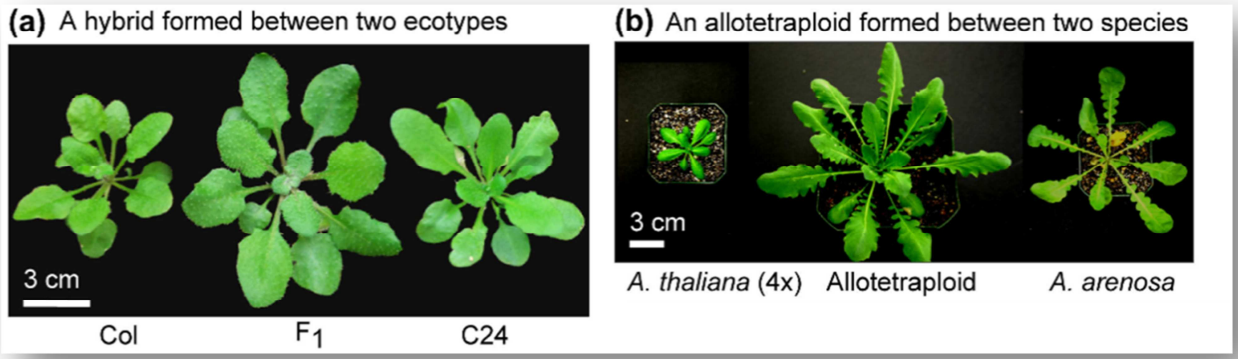


Figure 58. Phénomène d'hétérosis chez les hybrides et allotétraploïdes d'Arabidopsis.
(A) Plantules de l'hybride F1 (au milieu) et des lignées parentales *Arabidopsis thaliana Columbia* et *A. thaliana* C24. (B) Plantules de l'allotétraploïde (au milieu) et des lignées parentales *A. thaliana* Ler et *A. arenosa*. *Source : Chen et al. 2010*

Les spartines sont de bons modèles pour étudier séparément l'hybridation du doublement génomique puisque les hybrides et allopolyploïdes provenant des mêmes géniteurs sont disponibles dans la nature (Rousseau-Gueutin *et al.* 2016, Parisod *et al.* 2009a). Dans l'étude de Chelaifa *et al.* 2009, un effet différent a été observé entre hybride et allopolyploïde au niveau de l'expression des gènes. Une dominance maternelle est confirmée suite à l'hybridation, mais atténuée après la duplication du génome chez *S. anglica*. Le triticale est un hybride artificiel allopolyploïde entre le blé (dur ou tendre) et le seigle. Il illustre la réussite de l'hybridation interspécifique avec une meilleur croissance végétative, biomasse, et tolérance à des conditions défavorables (sols pauvres ou de stress hydrique). Fischer *et al.* 2010 ont montré que le gain en rendement pouvait s'élever à 12%. De récents articles démontrent l'implication d'une régulation épigénétique *via* l'expression de petits ARN dans le phénomène d'hétérosis des hybrides et de leur incompatibilité (Song *et al.* 2010, Chen *et al.*2007, 2010). Malgré l'importance de la polyploïdie et de la vigueur des hybrides en agriculture, ces mécanismes sont encore mal compris.

Sur cette base il serait pertinent de caractériser l'effet hétérosis chez le blé allopolyploïde naturel, le synthétique et chez le triticale, en lien avec la compartimentation des sous-génomes. Une large étude phénotypique de divers croisements (ainsi que les croisements réciproques ; interchangeant les mâles et les femelles) pourrait amener des réponses quant à l'apport des sous-génomes A, B et D du blé tendre dans ces phénomènes.

2.4. Comparaison avec le modèle animal et empreinte génomique

Basé sur la comparaison de l'histoire évolutive des espèces, le génome des animaux apparaît beaucoup plus stable et moins dynamique que celui des plantes (Murat *et al.* 2012). Alors que la polyploïdisation est très courante chez les plantes, le phénomène est rare dans le règne animal (certains poissons et amphibiens seraient des paléo-polyploïdes). Alors que les plantes gagnent en plasticité à la suite d'une duplication, elle est souvent délétère chez les animaux. Chez l'homme, 15 % des avortements spontanés

relèvent d'un événement de triploïdie. De plus, les cellules cancéreuses tendent à être polyploïdes (http://www.lamsade.dauphine.fr/). Non seulement les duplications sont rares chez les animaux, mais la teneur en éléments transposables est bien souvent inférieure aux plantes (45 % du génome chez l'homme et 80 % chez le blé). De plus, chez l'homme, les TE sont très anciens, datant d'environ 80 à 100 MYA, avec une distribution homogène le long des chromosomes. Tout ceci confère une certaine stabilité au génome des mammifères (Murat *et al.* 2012). Il est suggéré que les WGD ne sont plus possibles chez les animaux par la présence des chromosomes sexuels. La différenciation épigénétique du chromosome X rendrait irréalisable la polyploïdisation en raison d'une réduction impossible des gamètes post-polyploïdie. En effet, chez les espèces à reproduction sexuée lors de la polyploïdisation, la fusion des deux gamètes, assemble deux lots morphologiquement identiques de chromosomes fournis par chacun des parents pendant la fécondation.

Même s'ils sont rares, l'étude des hybrides animaux est intéressante car il est sans doute possible de faire émerger des hypothèses communes aux plantes et aux animaux. Les biologistes estiment que 10 % des animaux et 25 % des espèces végétales peuvent parfois se reproduire avec une autre espèce. Ainsi, les lions et les tigres par exemple, peuvent donner naissance aux ligers et tigrons. A travers ces hybrides on s'aperçoit que les génomes paternels et maternels ont un rôle différent. Ainsi, un liger est un croisement entre un lion mâle et un tigre femelle, tandis qu'un tigron est le croisement inverse (*cf.* Figure 59). Les traits des descendants varient considérablement en fonction de la direction du croisement parental. Les ligers sont les plus grands de tous les félins existants (*cf.* Figure 59), tandis que les tigrons sont de taille similaire aux parents.

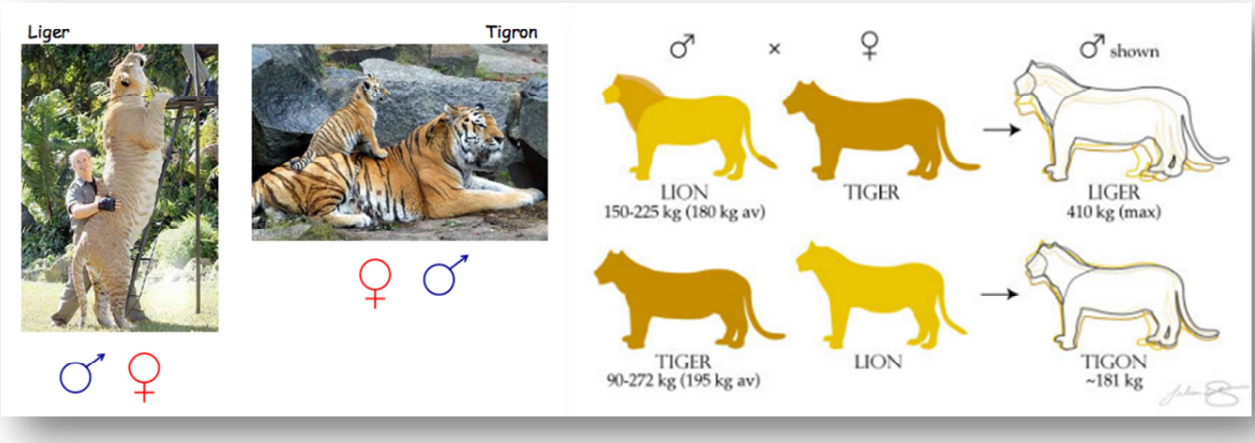


Figure 59. Illustration de l'hybridation interspécifique entre le lion et le tigre.
La photo de gauche représente un liger issu d'un croisement entre un lion mâle et un tigre femelle. La photo de droite représente un tigron issu d'un croisement entre un tigre mâle et un lionne. A droite sont représentés les caractéristiques (poids en kg) des lignées parentales et des hybrides. *Source : http://onlinelibrary.wiley.com/doi/10.1002/mrd.22074/epdf*

Ce phénomène est considéré comme le résultat des gènes soumis à empreinte, où les allèles paternels et maternels ne s'expriment pas de façon identique au cours du développement. Ainsi, chez les animaux

comme chez les plantes, il existe des gènes soumis à empreinte parentale. Les génomes parentaux sont fonctionnellement différents bien qu'ils contiennent une information génétique globalement identique ; ils sont porteurs d'une empreinte différentielle qui leur a été apposée avant la formation des pronoyaux dans le zygote. En effet, les empreintes (par exemple la méthylation des dinucléotides CpG avec répression de l'allèle) sont effacées avant la fécondation, et de nouvelles empreintes sont établies selon le sens du croisement. Les motifs de méthylations de l'ADN entrainent des modifications de la chromatine avec une acétylation et méthylation différentielle des histones des deux allèles. Chez les plantes à fleurs, une dizaine de gènes sous empreinte parentale ont été identifiés chez *Arabidopsis thaliana* et le maïs (Kohler *et al.*2010). Lors de la formation des gamètes, les allèles restent méthylés. Après la fécondation, les allèles maternels et paternels sont maintenus méthylés dans l'embryon, alors que l'allèle maternel est déméthylé dans l'albumen, grâce à l'activité ADN glycosylase de l'enzyme DEMETER, qui efface activement la méthylation de l'ADN. Ainsi selon l'origine parentale, l'allèle a un statut épigénétique différent. La perte répandue de méthylation de l'ADN par DEMETER est susceptible d'entraîner la réactivation transcriptionnelle de l'ensemble des transposons et séquences répétées, ce qui entraîne la production massive de siARN qui déclenche une méthylation asymétrique dans la cellule de l'œuf et l'albumen. Des études récentes (Kohler *et al.*2010, Kinoshita *et al.* 2007) tendent à démontrer que les gènes soumis à empreinte chez les plantes se trouvent proches de transposons ou séquences répétées, ce qui suggère que l'insertion de TE soit un prérequis à cet « imprinting ». Kinoshita *et al.* 2007 ont étudié le gène soumis à empreinte FWA (Late Flowering Phenotype). Ainsi, sa méthylation proviendrait d'un élément transposable de type SINE permettant le contrôle épigénétique du gène.

Sur cette base, la dynamique des TE régulés par les petits ARN lors de la fécondation amenant à une polyploïdisation peut avoir un rôle capital dans la cohabitation des sous-génomes d'un noyau. La caractérisation globale du choc génomique induit par cet évènement de duplication reste à approfondir. La dominance des sous-génomes est un processus lent et évolutif ; prend-elle naissance dès la fécondation ? Les allèles maternels ou paternels ont-ils un impact contrasté sur la mise en place de la compartimentation génomique? De nombreuses questions restent à éclaircir dans cette thématique. A ce niveau-là, l'étude et l'intégration de différents modèles (plantes, animaux, levure) sont intéressantes dans ce contexte en vue de faire émerger des nouvelles hypothèses en sciences fondamentales.

Chez le blé tendre *Triticum aestivum* (AABBDD), le génome chloroplastique provient d'*A. speltoides* (BB) tout comme pour les blés tétraploïdes (AABB) *Triticum turgidum* et *timopheevii* (Li. *et al.* 2015). Est-ce le hasard qui a fait que le génome sensible (du blé tendre et du blé dur) soit à chaque fois le génome cytoplasmique maternel ? Ou est-ce l'effet parental (c'est-à-dire l'origine paternelle ou maternelle lors de la fécondation) qui serait à l'origine de la sensibilité du sous-génome B du blé tendre depuis la fécondation? Sur cette base, l'étude de blés synthétiques bidirectionnels (Chapitre II, page 65) nous apprendra beaucoup sur le mécanisme de la dominance des sous-génomes et des effets maternels et paternels lors de la fécondation dans ce contexte. Chez les spartines les modifications structurales sont plus marquées sur le sous-génome maternel de S. *alterniflora* avec une altération de la méthylation des TE maternels (Parisod *et al.* 2010) au niveau de l'expression des gènes (dominance maternelle suite à fécondation ; Chelaifa *et al.* 2009).

CONCLUSION GLOBALE & PERSPECTIVES DE LA THESE

La polyploïdisation est un processus tout à fait naturel de doublement du matériel génétique qui joue un rôle majeur dans l'évolution des espèces et leurs spéciations. Lors de l'hybridation, une nouvelle espèce est produite résultant de deux sous-génomes différents (ou plus), enveloppés dans un même noyau. Ce génome post-polyploïdie doit alors orchestrer l'expression des gènes, les nouvelles interactions entre les facteurs de régulations des différents sous-génomes, et la réplication de l'ADN. Pour cela, il peut éliminer les gènes dupliqués ou, réprimer leurs expressions par des mécanismes génétiques ou épigénétiques ; ce qui mène *in fine* à la diploïdisation structurale, expressionnelle et fonctionnelle. Mes travaux de thèse ont montré que le blé tendre, pourtant considéré comme hexaploïde, est l'objet d'une diploïdisation en cours. A partir de quels éléments pouvons-nous dire qu'il est polyploïde ou diploïde ? Malgré leur origine hexaploïde (ou tétraploïde) les blés présentent un comportement diploïde à la méiose. Toutes les espèces de blés ne présentent que des bivalents à la méiose ; l'hérédité est dite de type disomique. Au niveau cytologique le locus Ph1 (*pairing homeologous*) bloque l'appariement des chromosomes homéologues chez le blé polyploïde pour permettre aux chromosomes homologues de s'apparier. Ainsi, ce gène limite l'aneuploïdie et favorise le maintien du comportement diploïde au niveau cytogénétique pour assurer la fertilité (ou descendance) de l'espèce. Il est envisageable que la perte des gènes au long court (induite par la diploïdisation structurale), lorsqu'elle parviendra à supprimer Ph1 (ainsi que les autres gènes de ce type), autorisera des remaniements chromosomiques (fusion, fission, aneuploïdie). Ainsi, elle installera la diploïdisation chromosomique en rompant ainsi, la seule trace symétrique de polyploïdie du blé tendre moderne.

Que la duplication soit ancienne ou récente, l'asymétrie observée et caractérisée dans ces travaux de thèse, peut avoir diverses origines. L'étude fine menée à la suite de mes travaux chez les blés hexaploïdes synthétiques apportera des réponses intéressantes sur les mécanismes à l'origine de la dominance des sous-génomes post-polyploïdie et notamment sur :

-l'expression des gènes des différents sous-génomes en réponse aux stress abiotiques ;

-l'expression des gènes des différents sous-génomes au regard de l'empreinte parentale ;

-le devenir des pré-miRNA et miRNA suite à un doublement génomique

-l'action des éléments transposables, leur rôle dans la pseudogénisation et l'inactivation des gènes ;

-le maintien de processus de régulations englobant les marques épigénétique sur plusieurs générations.

Le tout, en tenant compte de la compartimentation des sous-génomes A, B, D (héritée de la neo-hexaploïdisation), mais aussi en réponse à la paléoduplication commune à toutes les céréales, avec pour modèles les blés synthétiques principalement.

Pour conclure, doit-on parler du blé tendre en tant qu'espèce hexaploïde aujourd'hui dont le comportement méiotique est diploïde et dont nous montrons une quasi-complète diploïdisation structurale, expressionnelle et phénotypique ? Ces travaux permettent très certainement de poser la question d'une re-définition du concept « d'espèces polyploïdes » au regard des analyses génomiques qui peuvent être conduites aujourd'hui, comme cette thèse en est une illustration.

BIBLIOGRAPHIE

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, Butlin RK, Dieckmann U, Eroukhmanoff F, Grill A, Cahan SH, Hermansen JS, Hewitt G, Hudson AG, Jiggins C, Jones J, Keller B, Marczewski T, *et al.* Hybridization and speciation. J Evol Biol. 2013 Feb;26(2):229-46.

Abrouk M, Zhang R, Murat F, Li A, Pont C, Mao L, Salse J. Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. Plant Cell. 2012 May;24(5):1776-92.

Adams KL, Wendel JF. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 2005 Apr;8(2):135-41.

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien MA, Fukunaga K, Ricou M, Misset MT 2009. Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). Biological invasions 11: 1159–1173

Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Deal KR, Dubcovsky J, Gill BS, Gu YQ, Hadam J, Heo H, Huo N, Lazo GR, Luo MC, Ma YQ, Matthews DE, McGuire PE, Morrell PL, *et al.* Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. BMC Genomics. 2010 Dec 14;11:702.

Akhunov ED, Akhunova AR, Linkiewicz AM, Dubcovsky J, Hummel D, Lazo G, Chao S, Anderson OD, David J, Qi L, Echalier B, Gill BS, Miftahudin, Gustafson JP, La Rota M, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, *et al.* Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc Natl Acad Sci U S A. 2003 Sep 16;100(19):10836-41.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000 Dec 14;408(6814):796-815.

Baldrich P, Hsing YI, San Segundo B. Genome-Wide Analysis of Polycistronic MicroRNAs in Cultivated and Wild Rice. Genome Biol Evol. 2016 Apr 13;8(4):1104-14.

Baulcombe D. RNA silencing in plants. Nature. 2004 Sep 16;431(7006):356-63.

Blake NK, Lehfeldt BR, Lavin M, Talbert LE. Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome. 1999 Apr;42(2):351-60.

Blanc, G., & Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. The Plant Cell, 16(7), 1679–91.

Blatter RH, Jacomet S, Schlumbaum A. About the origin of European spelt (Triticum spelta L.): allelic differentiation of the HMW Glutenin B1-1 and A1-2 subunit genes. Theor Appl Genet. 2004 Jan;108(2):360-7.

Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J. The 'inner circle' of the cereal genomes. Curr Opin Plant Biol. 2009 Apr;12(2):119-25.

Bolser DM, Kerhornou A, Walts B, Kersey P. Triticeae resources in Ensembl Plants. Plant Cell Physiol. 2015 Jan;56(1):e3.

Borrill P, Adamski N, Uauy C. Genomics as the key to unlocking the polyploid potential of wheat. New Phytol. 2015 Dec;208(4):1008-22. doi: 10.1111/nph.13533. Epub 2015 Jun 24.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. 2012 Nov 29;491(7426):705-10.

C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998 Dec 11;282(5396):2012-8.

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Y, Xu Q, BianC, Zheng Z, Sun F, Liu W, Hsiao YY, Pan ZJ, Hsu CC, Yang YP, Hsu YC, Chuang YC,Dievart A, Dufayard JF, Xu X, Wang JY, Wang J, Xiao XJ, Zhao XM, Du R, Zhang GQ, Wang M, Su YY, Xie GC, Liu GH, Li LQ, Huang LQ, Luo YB, Chen HH, Van de Peer Y,Liu ZJ. The genome sequence of the orchid Phalaenopsis equestris. Nat Genet. 2015 Jan;47(1):65-72.

Campo S, Peris-Peris C, Siré C, Moreno AB, Donaire L, Zytnicki M, Notredame C, Llave C, San Segundo B. Identification of a novel microRNA (miRNA) from rice that targets an alternatively spliced transcript of the Nramp6 (Natural resistance-associated macrophage protein 6) gene involved in pathogen resistance. New Phytol. 2013 Jul;199(1):212-27.

Chagué V, Just J, Mestiri I, Balzergue S, Tanguy AM, Huneau C, Huteau V, Belcram H, Coriton O, Jahier J, Chalhoub B. Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol. 2010 Sep;187(4):1181-94.

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Corréa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, *et al.* Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science. 2014 Aug 22;345(6199):950-3.

Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 2010;11(12):R125.

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier MF, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). Plant Cell. 2005 Apr;17(4):1033-45.

Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L, Schmutz J, Yelick KA, Scholz U, Waugh R, Poland JA, Muehlbauer GJ, Stein N, Rokhsar DS. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 2015 Jan 31;16:26.

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B,Carter M, Huteau V, Coriton O, Appels R, Samain S, Chalhoub B. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics. 2008 Oct;180(2):1071-86.

Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, Wendel JF. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). Genetics. 2009 Jun;182(2):503-17.

Chelaifa H. Spéciation allopolyploïde et dynamique fonctionnelle du génome chez les Spartines. Université Rennes 1, 2010.

Chelaifa H, Mahé F, Ainouche M. Transcriptome divergence between the hexaploid salt-marsh sister species Spartina maritima and Spartina alterniflora (Poaceae). Mol Ecol. 2010 May;19(10):2050-63.

Chen ZJ. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci. 2010 Feb;15(2):57-71.

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X. Deciphering the diploid ancestral genome of the Mesohexaploid Brassica rapa. Plant Cell. 2013 May;25(5):1541-54.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One. 2012;7(5):e36442.

Cheng Z, Buell CR, Wing RA, Gu M, Jiang J. Toward a cytological characterization of the rice genome. Genome Res. 2001 Dec;11(12):2133-41.

Chimungu JG, Brown KM, Lynch JP. Large root cortical cell size improves drought tolerance in maize. Plant Physiol. 2014 Dec;166(4):2166-78.

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, *et al.* Structural and functional partitioning of bread wheat chromosome 3B. Science. 2014 Jul 18;345(6194):1249721.

Cifuentes M, Grandont L, Moore G, Chèvre AM, Jenczewski E. Genetic regulation of meiosis in polyploid species: new insights into an old question. New Phytol. 2010 Apr;186(1):29-36.

Dahm R. Friedrich Miescher and the discovery of DNA. Dev Biol. 2005 Feb 15;278(2):274-88.

Darwin C. 1859. On the origin of species by natural selection. London: Murray

Devos KM, Atkinson MD, Chinoy CN, Liu CJ, Gale MD. RFLP-based genetic map of the homoeologous group 3 chromosomes of wheat and rye. Theor Appl Genet. 1992 May;83(8):931-9.

Dibari B, Murat F, Chosson A, Gautier V, Poncet C, Lecomte P, Mercier I, Bergès H, Pont C, Blanco A, Salse J. Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. BMC Genomics. 2012 Jun 6;13:221.

Dobrovolskaya O, Pont C, Sibout R, Martinek P, Badaeva E, Murat F, Chosson A, Watanabe N, Prat E, Gautier N, Gautier V, Poncet C, Orlov YL, Krasnikov AA, Bergès H, Salina E, Laikova L, Salse J. FRIZZY PANICLE drives supernumerary spikelets in bread wheat. Plant Physiol. 2015 Jan;167(1):189-99.

Donald, C. M. 1968. The breeding of crop ideotypes. Euphytica 17:385–403.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet. 2008;42:443-61.

Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science. 2007 Jun 29;316(5833):1862-6.

Duggan B.L., Richards R.A., Van Herwaarden A.F., Fettell N.A. (2006) Agronomic evaluation of a tiller inhibition gene (tin) in wheat. I. Effect on yield, yiel components, and grain protein. Australian Journal of Agricultural Research 56(2) 169-178.

Elliott, Plant breeding and cytogenetics, New York : McGraw-Hill, 1958.

Emmanuel PJ. Polymerase chain reaction from bench to bedside. Applications for infectious disease. J Fla Med Assoc. 1993 Sep;80(9):627-30.

Endo, T. R., and B. S. Gill, 1996 The deletion stocks of common wheat. J. Hered. 87: 295-307.

Feldman M. Cytogenetic Activity and Mode of Action of the Pairing Homoeologous (Phi) Gene of Wheat. Crop Science 33(5) · January 1993.

Feldman M, Levy AA, Fahima T, Korol A. Genomic asymmetry in allopolyploid plants: wheat as a model. J Exp Bot. 2012 Sep;63(14):5045-59.

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, Jia P, Zhang Y, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Lu Y, Zhang LS, *et al.* Sequence and analysis of rice chromosome 4. Nature. 2002 Nov 21;420(6913):316-20.

Fischer S, H. P. Maurer, T. Würschum, J. Möhring, H.-P. Piepho, C. C. Schön, E.-M. Thiemt, B. S. Dhillon, E.A. Weissmann, A. E. Melchinger, and J. C. Reif. Development of Heterotic Groups in Triticale. crop science, vol. 50, march–april 2010

Flagel LE, Wendel JF. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol. 2010 Apr;186(1):184-93.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496-512.

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol. 2012 Apr;15(2):131-9.

Friebe, B., and B. S. Gill, 1994 C-band polymorphism and structural rearrangements detected in common wheat (Triticum aestivum). Euphytica 78: 1-5.

Fu D, Szucs P, Yan L, Helguera M, Skinner JS, von Zitzewitz J, Hayes PM, Dubcovsky J. Large deletions within the first intron in VRN-1 are associated with spring growth habit in barley and wheat. Mol Genet Genomics. 2005 Mar;273(1):54-65. Epub 2005 Feb 3. Erratum in: Mol Genet Genomics. 2005 Nov;274(4):442-3.

Galili G, Feldman M. Genetic control of endosperm proteins in wheat : 2. Variation in high molecular weight glutenin and gliadin subunits of Triticum aestivum. Theor Appl Genet. 1983 Jul;66(1):77-86.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 2014 Feb;31(2):448-54.

Gill, BS, Friebe, B, Endo, TR. Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (Triticum aestivum). Genome, 34: 830

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature. 2006 Feb 9;439(7077):749-52.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science. 2002 Apr 5;296(5565):92-100.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, *et al.* A draft sequence of the Neandertal genome. Science. 2010 May 7;328(5979):710-22.

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, Charles M, Anderson OD, Chalhoub B. Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics. 2006 Nov;174(3):1493-504.

Guo X, Zhang Z, Gerstein MB, Zheng D. Small RNAs originated from pseudogenes: cis- or trans-acting? PLoS Comput Biol. 2009 Jul;5(7):e1000449.

He P, Friebe BR, Gill BS, Zhou JM. Allopolyploidy alters gene expression in the highly stable hexaploid wheat. Plant Mol Biol. 2003 May;52(2):401-14.

Heat Priming Induces Trans-generational Tolerance to High Temperature Stress in Wheat. Wang X, Xin C, Cai J, Zhou Q, Dai T, Cao W, Jiang D. Frontiers in Plant Science. 1/01/01 00:00; 7: 501 PMC [article] PMCID: PMC4830833, PMID: 27148324,

Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009 Aug;19(8):1419-28.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45.

International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science. 2014 Jul 18;345(6194):1251788.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, *et al.* Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013 Apr 4;496(7443):91-5.

Jiao Y, Paterson AH. Polyploidy-associated genome modifications during land plant evolution. Philos Trans R Soc Lond B Biol Sci. 2014 Aug 5;369(1648).

Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL; IWGS Consortium, Sharpe AG, Sidebottom CH, Hall N, Toomajian C, Close, T, Dubcovsky J, Akhunova A, Talbert L, Bansal UK, Bariana HS, Hayden MJ, Pozniak, C, Jeddeloh JA, Hall A, Akhunov E. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol. 2015 Feb, 26;16:48.

Kantar M, Unver T, Budak H. Regulation of barley miRNAs upon dehydration stress correlated with target gene expression. Funct Integr Genomics. 2010 Nov;10(4):493-507.

Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics. 2002 Apr;160(4):1651-9.

Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet. 2003 Jan;33(1):102-6.

Kebrom TH, Spielmeyer W, Finnegan EJ. Grasses provide new insights into regulation of shoot branching. Trends Plant Sci. 2013 Jan;18(1):41-8.

Kilian B, Ozkan H, Deusch O, Effgen S, Brandolini A, Kohl J, Martin W, Salamini F. Independent wheat B and G genome origins in outcrossing Aegilops kilprogenitor haplotypes. Mol Biol Evol. 2007 Jan;24(1):217-27.

Kim JS, Islam-Faridi MN, Klein PE, Stelly DM, Price HJ, Klein RR, Mullet JE. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics. 2005 Dec;171(4):1963-76.

Kinoshita Y, Saze H, Kinoshita T, Miura A, Soppe WJ, Koornneef M, Kakutani T. Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. Plant J. 2007 Jan;49(1):38-45.

Kruszka K, Pacak A, Swida-Barteczka A, Stefaniak AK, Kaja E, Sierocka I, Karlowski W, Jarmolowski A, Szweykowska-Kulinska Z. Developmentally regulated expression and complex processing of barley pri-microRNAs. BMC Genomics. 2013 Jan 16;14:34.

Kuraparthy V, Sood S, Dhaliwal HS, Chhuneja P, Gill BS. Identification and mapping of a tiller inhibition gene (tin3) in wheat. Theor Appl Genet. 2007 Jan;114(2):285-94.

Lamarck J.B. 1809. La philosophie zoologique. Paris : Dentu et l'auteur, 2 vol., 856p. + 950p.

Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990 Mar;124(3):743-56.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9:559.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923.

Lee EJ, Pei L, Srivastava G, Joshi T, Kushwaha G, Choi JH, Robertson KD, Wang X, Colbourne JK, Zhang L, Schroth GP, Xu D, Zhang K, Shi H. Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res. 2011 Oct;39(19):e127. doi: 10.1093/nar/gkr598.

Levy AA, Feldman M. The impact of polyploidy on grass genome evolution. Plant Physiol. 2002 Dec;130(4):1587-93.

Levy AA, Galili G, Feldman M. 1988. Polymorphism and genetic control of high molecular weight glutenin subunits in wild tetraploid wheat Triticum turgidum var. dicoccoides. Heredity61, 63-72

Li A, Liu D, Wu J, Zhao X, Hao M, Geng S, Yan J, Jiang X, Zhang L, Wu J, Yin L, Zhang R, Wu L, Zheng Y, Mao L. mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat. Plant Cell. 2014 May 16;26(5):1878-1900.

Li LF, Liu B, Olsen KM, Wendel JF. A re-evaluation of the homoploid hybrid origin of Aegilops tauschii, the donor of the wheat D-subgenome. New Phytol. 2015a Oct;208(1):4-8.

Li LF, Liu B, Olsen KM, Wendel JF. Multiple rounds of ancient and recent hybridizations have occurred within the Aegilops-Triticum complex. New Phytol. 2015b Oct;208(1):11-2.

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, *et al.* Draft genome of the wheat A-genome progenitor Triticum urartu. Nature. 2013 Apr 4;496(7443):87-90.

Lippman Z, Martienssen R. The role of RNA interference in heterochromatic silencing. Nature. 2004 Sep 16;431(7006):364-70.

Loukoianov A, Yan L, Blechl A, Sanchez A, Dubcovsky J. Regulation of VRN-1 vernalization genes in normal and transgenic polyploid wheat. Plant Physiol. 2005 Aug;138(4):2364-73. Epub 2005 Jul 29.

Lucas SJ, Budak H. Sorting the wheat from the chaff: identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. PLoS One. 2012;7(7):e40859.

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, Martis MM, Mayer KF, Sehgal SK, *et al.* 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 U S A. 2013 May 7;110(19):7940-5.

McClintock B. 1984. The significance of responses of the genome to challenge. Science 226: 792– 801.

Madlung A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity (Edinb). 2013 Feb;110(2):99-104.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M; International Wheat Genome Sequencing Consortium,, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF, Olsen OA. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 2014 Jul 18;345(6194):1250092.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., & Otto, S. P. (2011). Recently formed polyploid plants diversify at lower rates. Science (New York, N.Y.), 333(6047), 1257.

Mehmet S, Characterisation of dna from archaeological wheat (Triticum L.) seeds from anatolia. A thesis submitted to the graduate school of natural and applied sciences of the middle east technical university. 2003

Merchan F, Boualem A, Crespi M, Frugier F. Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins. Genome Biol. 2009;10(12):R136.

Mestiri I, Changements génétiques et épigénétiques en relation avec le comportement méiotique chez les allopolyploïdes de blé (genres Aegilopset Triticum). Présentée pour obtenir le grade de Docteur en sciences de l'université d'Évry-Val d'Essonne. 2010.

Michael S. Barker, Gregory J. Baute, and Shao-Lun Liu; J.F. Wendel et al., Duplications and Turnover in Plant Genomes, Plant Genome Diversity Volume 1

Mitchell JH, Rebetzke GJ, Chapman SC, Fukai S. Evaluation of reduced-tillering (tin) wheat lines in managed, terminal water deficit environments. J Exp Bot. 2013 Aug;64(11):3439-51.

Moeller C, Evers JB, Rebetzke G. Canopy architectural and physiological characterization of near-isogenic wheat lines differing in the tiller inhibition gene tin. Front Plant Sci. 2014;5:617.

Moore, G. (2002). Meiosis in allopolyploids -- the importance of "Teflon" chromosomes. Trends in Genetics :TIG, 18(9), 456–63.

Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution. Grasses, line up and form a circle. Curr Biol. 1995 Jul 1;5(7):737-9.

Mosher RA, Melnyk CW, Kelly KA, Dunn RM, Studholme DJ, Baulcombe DC. Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. Nature. 2009 Jul 9;460(7252):283-6.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Roest Crollius H, Salse J. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 2015 Dec 10;16:262.

Murat F, Van de Peer Y, Salse J. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome Biol Evol. 2012;4(9):917-28.

Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res. 2010 Nov;20(11):1545-57.

Murat F, Zhang R, Guizard S, Flores R, Armero A, Pont C, Steinbach D, Quesneville H, Cooke R, Salse J. Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. Genome Biol Evol. 2014 Jan;6(1):12-33.

Murat F, Zhang R, Guizard S, Gavranovi? H, Flores R, Steinbach D, Quesneville H, Tannier E, Salse J. Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops.Genome Biol Evol. 2015 Jan 29;7(3):735-49.

Müntzing, A. (1936). The Evolutionary Significance of Autopolyploidy. Hereditas, 21(2-3), 363–378.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, *et al.* A whole-genome assembly of Drosophila. Science. 2000 Mar 24;287(5461):2196-204.

Nelissen H, Moloney M, Inzé D. Translational research: from pot to plot. Plant Biotechnol J. 2014 Apr;12(3):277-85.

Pandey R, Joshi G, Bhardwaj AR, Agarwal M, Katiyar-Agarwal S. A comprehensive genome-wide study on tissue-specific and abiotic stress-specific miRNAs in Triticum aestivum. PLoS One. 2014;9(4):e95800.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien MA, Ainouche M. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol. 2009 Dec;184(4):1003-15.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 2010 Apr;186(1):37-45.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009).The Sorghum bicolor genome and the diversification of grasses. Nature, 457(7229), 551–6.

Paterson, A. H., Bowers, J. E., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America, 101(26), 9903–8.

Pearce S, Saville R, Vaughan SP, Chandler PM, Wilhelm EP, Sparks CA, Al-Kaff N, Korolev A, Boulton MI, Phillips AL, Hedden P, Nicholson P, Thomas SG. Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. Plant Physiol. 2011 Dec;157(4):1820-31.

Pellegrineschi A, Noguera LM, Skovmand B, Brito RM, Velazquez L, Salgado MM, Hernandez R, Warburton M, Hoisington D. Identification of highly transformable wheat genotypes for mass production of fertile transgenic plants. Genome. 2002 Apr;45(2):421-30. PubMed [citation] PMID: 11962639

Peng J, Korol AB, Fahima T, Röder MS, Ronin YI, Li YC, Nevo E. Molecular genetic maps in wild emmer wheat, Triticum dicoccoides: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Res. 2000 Oct;10(10):1509-31.

Peng J, Ronin Y, Fahima T, Röder MS, Li Y, Nevo E, Korol A. Domestication quantitative trait loci in Triticum dicoccoides, the progenitor of wheat. Proc Natl Acad Sci U S A. 2003 Mar 4;100(5):2489-94.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR; International Wheat Genome Sequencing Consortium, Mayer KF, Olsen OA. Genome interplay in the grain transcriptome of hexaploid bread wheat. Science. 2014 Jul 18;345(6194):1250091.

Piperno DR, Sues HD. 2005. Dinosaurs dined on grass. Science, 310: 1126-1128.

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006 Jan 20;311(5759):392-4.

Pont C, Murat F, Guizard S, Flores R, Foucrier S, Bidet Y, Quraishi UM, Alaux M, Doležel J, Fahima T, Budak H, Keller B, Salvi S, Maccaferri M, Steinbach D, Feuillet C, Quesneville H, Salse J. Wheat synténome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J. 2013 Dec;76(6):1030-44.

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorák J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, *et al.* A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics. 2004 Oct;168(2):701-12.

Quraishi UM, Abrouk M, Murat F, Pont C, Foucrier S, Desmaizieres G, Confolent C, Rivière N, Charmet G, Paux E, Murigneux A, Guerreiro L, Lafarge S, Le Gouis J, Feuillet C, Salse J. Cross-genome map based dissection of a nitrogen use efficiency ortho-metaQTL in bread wheat unravels concerted cereal genome evolution. Plant J. 2011 Mar;65(5):745-56.

Quraishi UM, Murat F, Abrouk M, Pont C, Confolent C, Oury FX, Ward J, Boros D, Gebruers K, Delcour JA, Courtin CM, Bedo Z, Saulnier L, Guillon F, Balzergue S, Shewry PR, Feuillet C, Charmet G, Salse J. Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (Triticum aestivum L.). Funct Integr Genomics. 2011 Mar;11(1):71-83.

Ramsey J, Schemske DW (2002) Neopolyploidy in flowering plants. Ann Rev Ecol Systemat 33: 589–639.

Reduced-tillering wheat lines maintain kernel wight in dry environments. J.H Mitchell, SC. Chapman, GJ Rebetzke and Shu Fukai. 2006 13th Australian agronomy conference.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. Mol Biol Evol. 2015 Apr;32(4):1063-71.

Renny-Byfield, S., & Wendel, J. F. (2014). Doubling down on genomes: Polyploidy and crop plants. American Journal of Botany, 101(10), 1711–1725.

Richards RA. (1988) A tiller inhibitor gene in wheat and its effect on plant-growth. Aust J Agric Res 39: 749-757

Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW. A microsatellite map of wheat. Genetics. 1998 Aug;149(4):2007-23.

Rousseau-Gueutin M, Bellot S, Martin GE, Boutte J, Chelaifa H, Lima O, Michon-Coudouel S, Naquin D, Salmon A, Ainouche K, Ainouche M. The chloroplast genome of the hexaploid Spartina maritima (Poaceae, Chloridoideae): Comparative analyses and molecular dating. Mol Phylogenet Evol. 2015 Dec;93:5-16.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc Natl Acad Sci U S A. 2009 Sep 1;106(35):14908-13.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell. 2008a Jan;20(1):11-24.

Salse J, Chagué V, Bolot S, Magdelenat G, Huneau C, Pont C, Belcram H, Couloux, A, Gardais S, Evrard A, Segurens B, Charles M, Ravel C, Samain S, Charmet G, Boudet N, Chalhoub B. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. BMC Genomics. 2008b Nov 25;9:555.

Sánchez-Morán E, Benavente E, Orellana J. Analysis of karyotypic stability of homoeologous-pairing (ph) mutants in allopolyploid wheats. Chromosoma. 2001 Sep;110(5):371-7.

Sandve SR, Marcussen T, Mayer K, Jakobsen KS, Heier L, Steuernagel B, Wulff BB, Olsen OA. Chloroplast phylogeny of Triticum/Aegilops species is not incongruent with an ancient homoploid hybrid origin of the ancestor of the bread wheat D-genome. New Phytol. 2015 Oct;208(1):9-10.

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, *et al.* The genome sequence and structure of rice chromosome 1. Nature. 2002 Nov 21;420(6913):312-6.

Schnable JC, Freeling M, Lyons E. Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol. 2012;4(3):265-77.

Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 2011 Mar 8;108(10):4069-74.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science (New York, N.Y.), 326(5956), 1112–5

Schranz ME, Osborn TC. De novo variation in life-history traits and responses to growth conditions of resynthesized polyploid Brassica napus (Brassicaceae). Am J Bot. 2004 Feb;91(2):174-83.

Sears, E. R., 1954 The Aneuploids of Common Wheat. Bull. 572, University of Missouri Agricultural Experiment Station, Columbia, MO

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. The Plant Cell. 2001/08/01 00:00; 13(8): 1749-1760

Shapiro, Beth, Hofreiter, Michael (Eds.). Ancient DNA: Methods and Protocols, Methods in Molecular Biology, Vol. 840. Springer protocols, 2012, 241p.

Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijó JA, Martienssen RA. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009 Feb 6;136(3):461-72.

Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then…and now: Stebbins revisited. Am J Bot. 2014 Jul 20;101(7):1057-1078.

Soltis PS, Soltis DE. The role of hybridization in plant speciation. Annu Rev Plant Biol. 2009;60:561-88.

Song Q, Chen ZJ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol. 2015 Apr;24:101-9.

Spielmeyer W, Richards RA. Comparative mapping of wheat chromosome 1AS which contains the tiller inhibition gene (tin) with rice chromosome 5S. Theor Appl Genet. 2004 Oct;109(6):1303-10.

Suenaga K, Khairallah M, William HM, Hoisington DA. A new intervarietal linkage map and its application for quantitative trait locus analysis of "gigas" features in bread wheat. Genome. 2005 Feb;48(1):65-75.

Sugano SS, Shirakawa M, Takagi J, Matsuda Y, Shimada T, Hara-Nishimura I, Kohchi T. CRISPR/Cas9-mediated targeted mutagenesis in the liverwort Marchantia polymorpha L. Plant Cell Physiol. 2014 Mar;55(3):475-81.

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu JK. Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol. 2008 Feb 29;8:25.

Talbert LE, Blake NK, Storlie EW, Lavin M. Variability in wheat based on low-copy DNA sequence comparisons. Genome. 1995 Oct;38(5):951-7.

Tavakol E, Okagaki R, Verderio G, Shariati J. V, Hussien A, Bilgic H, Scanlon MJ, Todt NR, Close TJ, Druka A, Waugh R, Steuernagel B, Ariyadasa R, Himmelbach A, Stein N, Muehlbauer GJ, Rossini L.The Barley Uniculme4 Gene Encodes a BLADE-ON-PETIOLE-Like Protein That Controls Tillering and Leaf Patterning. Plant Physiology. 2015/05/01 00:00; 168(1): 164-174 PMC [article] PMCID: PMC4424007, PMID: 25818702,

The Economist, Feb. 26th 2011, The 9 billion-people question. A special report on feeding the world.

Thomas BC, Pedersen B, 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(7):934-46.

Throude M, Bolot S, Bosio M, Pont C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, Murigneux A, Charmet G, Perez P, Salse J. Structure and expression analysis of rice paleo duplications. Nucleic Acids Res. 2009 Mar;37(4):1248-59.

Tittel-Elmer M, Bucher E, Broger L, Mathieu O, Paszkowski J, Vaillant I. Stress-induced activation of heterochromatic transcription. PLoS Genet. 2010 Oct 28;6(10):e1001175.

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nature Reviews. Genetics, 10(10), 725–32.

Wan Y, Poole RL, Huttly AK, Toscano-Underwood C, Feeney K, Welham S, Gooding MJ, Mills C, Edwards KJ, Shewry PR, Mitchell RA. Transcriptome analysis of grain development in hexaploid wheat. BMC Genomics. 2008 Mar 6;9:121.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, *et al.* The genome of the mesopolyploid crop species Brassica rapa. Nat Genet. 2011 Aug 28;43(10):1035-9.

Wang Z, Liu Y, Shi H, Mo H, Wu F, Lin Y, Gao S, Wang J, Wei Y, Liu C, Zheng Y. Identification and validation of novel low-tiller number QTL in common wheat. Theor Appl Genet. 2016 Mar;129(3):603-12.

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J; International Wheat Genome Sequencing Consortium, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. Plant Biotechnol J. 2014 Aug;12(6):787-96.

Wright M, Dawson J, Dunder E, Suttie J, Reed J, Kramer C, Chang Y, Novitzky R, Wang H, Artim-Moore L. 2001 Efficient biolistic transformation of maize (Zea mays L.) and wheat (Triticum aestivum L.) using the phosphomannose isomerase gene, pmi, as the selectable marker Plant Cell Reports 20 429–436.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol. 2010 Jun29;8(6):e1000409.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., & Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences of the United States of America, 106(33), 13875–9.

Xia J, Han L, Zhao Z. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC Genomics. 2012;13 Suppl 8:S7.

Yang L, Takuno S, Waters ER, Gaut BS. Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation. Mol Biol Evol. 2011 Mar;28(3):1193-203.

Yoo MJ, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb). 2013 Feb;110(2):171-80.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science. 2002 Apr 5;296(5565):79-92.

Zhang H, Bian Y, Gou X, Zhu B, Xu C, Qi B, Li N, Rustgi S, Zhou H, Han F, Jiang J, von Wettstein D, Liu B. Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. Proc Natl Acad Sci U S A. 2013 Feb 26;110(9):3447-52.

Zhang Z, Belcram H, Gornicki P, Charles M, Just J, Huneau C, Magdelenat G, Couloux A, Samain S, Gill BS, Rasmussen JB, Barbe V, Faris JD, Chalhoub B. Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci U S A. 2011 Nov 15;108(46):18737-42.

ANNEXE 1

El Baidouri M, Murat F, Veyssiere M, Molinier M, Pont C*, Salse J*. Reconciliating the evolutionary origin of bread wheat (Triticum aestivum) (Chapitre 1)

New Phytologist

Research

Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*)

Moaine El Baidouri[1], Florent Murat[1], Maeva Veyssiere[1], Mélanie Molinier[1], Raphael Flores[2], Laura Burlot[2], Michael Alaux[2], Hadi Quesneville[2], Caroline Pont[1] and Jérôme Salse[1]

[1]INRA/UBP UMR 1095 GDEC (Genetics, Diversity and Ecophysiology of Cereals), 5 chemin de Beaulieu, Clermont Ferrand 63100, France; [2]INRA UR1164 URGI (Research Unit in Genomics-Info), Université Paris-Saclay, Versailles 78026, France

Authors for correspondence:
Jérôme Salse
Tel: +33 0 473624380
Email: jsalse@clermont.inra.fr

Caroline Pont
Tel: +33 0 473624300
Email: cpont@clermont.inra.fr

Received: *17 February 2016*
Accepted: *18 June 2016*

New Phytologist (2016)
doi: 10.1111/nph.14113

Key words: ancestor, evolution, origin, polyploidization, wheat.

Summary

- The origin of bread wheat (*Triticum aestivum*; AABBDD) has been a subject of controversy and of intense debate in the scientific community over the last few decades. In 2015, three articles published in *New Phytologist* discussed the origin of hexaploid bread wheat (AABBDD) from the diploid progenitors *Triticum urartu* (AA), a relative of *Aegilops speltoides* (BB) and *Triticum tauschii* (DD).
- Access to new genomic resources since 2013 has offered the opportunity to gain novel insights into the paleohistory of modern bread wheat, allowing characterization of its origin from its diploid progenitors at unprecedented resolution.
- We propose a reconciled evolutionary scenario for the modern bread wheat genome based on the complementary investigation of transposable element and mutation dynamics between diploid, tetraploid and hexaploid wheat.
- In this scenario, the structural asymmetry observed between the A, B and D subgenomes in hexaploid bread wheat derives from the cumulative effect of diploid progenitor divergence, the hybrid origin of the D subgenome, and subgenome partitioning following the polyploidization events.

Introduction

Bread wheat (*Triticum aestivum*) evolved through two polyploidization events between *Triticum urartu* (AA genome) and an *Aegilops speltoides*-related species (BB genome) 0.5 million yr ago (hereafter Ma), forming *Triticum turgidum* ssp. *diccocoides*, and between *Triticum turgidum* ssp. *durum* (AABB genome) and *Aegilops tauschii* (DD genome) 10 000 yr ago, forming the modern hexaploid bread wheat (AABBDD) genome (Feldman *et al.*, 1995; Huang *et al.*, 2002). Recently available wheat genomic resources offered the opportunity to gain novel insights into the origin of wheat with the release of the genome shotgun sequences of hexaploid and tetraploid wheat (IWGSC, 2014) as well as diploid progenitors (Jia *et al.*, 2013; Ling *et al.*, 2013; Luo *et al.*, 2013).

From these resources, Marcussen *et al.* (2014), confirmed in Sandve *et al.* (2015), estimated the phylogenetic history of the A, B and D subgenomes from 2269 gene trees involving A, B and D homoeologs conserved between the hexaploid wheat subgenomes, among which 275 trees include orthologous sequences from five diploid relatives (*T. urartu*, *A. speltoides*, *A. tauschii*, *Triticum monococcum* and *Aegilops sharonensis*). The authors reported that the two tree typologies A(B/D) and B(A/D) were twice as abundant as D(A/B). This gene-based phylogenetic approach then revealed that the A and B subgenomes are more closely related individually to the D subgenome than to each other. The authors then proposed that the D genome originated from a homoploid ancestor derived from the hybridization of the A and B diploid progenitors 5 Ma.

Li *et al.* (2015a), confirmed in Li *et al.* (2015b), re-evaluated the origin of hexaploid bread wheat based on the phylogenomic investigation of 20 chloroplast genomes, which are maternally inherited in this species complex. The authors argued that, in Marcussen *et al.*'s (2014) scenario of a homoploid origin of the D subgenome, *A. tauschii* would be expected to share the chloroplast genome of one (the maternal) of the two progenitors (either *T. urartu* or *A. speltoides*). Instead, the authors reported a nested topology of the *A. taushii* chloroplast genome. Taking into account not only the A, B and D progenitor genomes but also the M, N, T, U and C (referred to as S) diploid relatives within this species complex, the authors reported that the chloroplast genome of *A. taushii* (D) is more closely related to other D and S genomes than to the genomes of *A. speltoides* (S) and *T. urartu* (A). Li *et al.* (2015a,b) then hypothesized that the origin of the D (*A. tauschii*) genome may be more complex (additional hybridization events to be considered) than suggested initially by Marcussen *et al.* (2014).

In addition to previous investigations of the evolutionary history of the hexaploid wheat D subgenome, the origin of the B subgenome has also been the subject of intense debate. Several phylogenetic studies have tried to identify the progenitor of the B

genome of polyploid wheat based on cytology (Zohary & Feldman, 1962), nuclear and mitochondrial DNA sequences (Dvorak *et al.*, 1989; Dvorak & Zhang, 1990; Terachi *et al.*, 1990) and chromosome rearrangement studies (Feldman, 1966a, b; Hutchinson *et al.*, 1982; Gill & Chen, 1987; Naranjo *et al.*, 1987; Naranjo, 1990; Jiang & Gill, 1994; Devos *et al.*, 1995; Maestra & Naranjo, 1999). Molecular comparisons at the whole-genome level using germplasm collections have shown that the B subgenome from hexaploid wheat could be related to several *A. speltoides* lines but not to other species of the *Sitopsis* section (Salina *et al.*, 2006; Kilian *et al.*, 2007). At the Storage Protein Activator (*SPA*) locus (Salse *et al.*, 2008), close relationships between *A. speltoides* and the hexaploid B subgenome have been reported based on both coding and noncoding sequence comparisons, but with lower conservation compared with the A subgenome and its *T. urartu* progenitor at the *putative ATP binding cassette* (*ABC*) *transporter gene* (*PSR920*) locus (Dvorak & Akhunov, 2005; Dvorak *et al.*, 2006). Taken together, the findings of these studies suggest two hypotheses, the first being that the progenitor of the B genome is a unique and ancient *Aegilops* species that remains unknown (i.e. monophyletic origin and ancestor closely related to *A. speltoides* from the *Sitopsis* section), and the second being that the B genome resulted from the introgression of several parental *Aegilops* species (i.e. polyphyletic origin) from the *Sitopsis* section that need to be identified.

Several research groups have suggested the hypothesis of a single ancient hybridization event (Sandve *et al.*, 2015) or nested rounds of hybridization events (Li *et al.*, 2015a,b) at the origin of the wheat D subgenomes; and several studies also proposed two possible origins of the B subgenome (i.e. either mono- or polyphyletic). In the current study, we investigated the evolutionary dynamics of gene-based transposable elements (TEs) and mutations (single nucleotide mutations between homoeologs, homoeoSNPs) between the A, B and D subgenomes as well as between hexaploid, tetraploid and diploid wheat to redraw the origin of the modern bread wheat genome.

Materials and Methods

Wheat genome sequences

The genome sequences (as reported in IWGSC, 2014) used to reveal the origin of the modern bread wheat genome correspond to *Triticum aestivum* (AABBDD; 99 386 genes), *Triticum durum* (AABB; 91 097 genes), *Triticum urartu* (AA; 53 056 genes), *Aegilops speltoides* (BB; 62 258 genes) and *Aegilops tauschii* (DD; 50 264 genes).

Wheat syntenome construction

The ancestral grass genome (ancestral grass karyotype (AGK)) as reported in Murat *et al.* (2014) was used, with 58 933 ordered ancestral genes on 12 ancestral chromosomes based on synteny relationships between the *Oryza sativa* (rice, IRGSP, 2005), *Brachypodium distachyon* (Brachypodium, IBI, 2010) and *Sorghum bicolor* (sorghum, Paterson *et al.*, 2009) genomes. The BlastN alignment of 40 267 mapped markers from the wheat consensus single nucleoide polymorphism (SNP) map published by Wang *et al.* (2014) and AGK genes yielded orthologs between these two resources. Using the DRIMM-Synteny tool (Pham & Pevzner, 2010), we built synteny groups allowing the identification of ancestral regions, ancestral gene content and finally the order of wheat genes on the consensus map (21 chromosomes) (Pont *et al.*, 2013). Following this method, we ordered 62 135 wheat sequence scaffolds (from IWGSC, 2014) containing 72 900 genes along the 21 chromosomes of the bread wheat genome, which was termed the bread wheat syntenome (available at http://urgi.versailles.inra.fr/synteny-wheat).

Identification of homoeologous triplets

A BlastN all-against-all search was performed using the 99 386 predicted wheat genes (Borrill *et al.*, 2015) in order to define A, B and D homoeologs. Genes sharing a cumulative identity percentage (CIP) of > 90% and a cumulative alignment length percentage (CALP) of at least 30% (Salse *et al.*, 2009) were grouped in the same cluster using the Markov cluster (mcl) algorithm (http://micans.org/mcl/). Clusters containing strictly three genes belonging to the A, B and D subgenomes of the same chromosomal group were considered as robust homoeologous genes (8671 homoeologous triplets were identified).

TE insertional dynamics analysis

The 8671 homoeologous gene triplets were automatically scanned using Mummer (http://mummer.sourceforge.net/manual/) in order to detect sequence homology breakpoints between homoeologs that are potentially caused by TE insertions. Putative shared TE insertions were then manually checked using Dotter (http://sonnhammer.sbc.su.se/Dotter.html) and only breakpoints corresponding to the exact TE boundaries were retained for further analysis. Empty TE sites in homoeologs can be a hallmark of either absence of the insertion (demonstrated by the absence of target site duplication (TSD)) or the excision of the considered element (demonstrated by the presence of at least remnants of TSD), as the investigated class II elements transpose via a 'cut and paste' mechanism. For these reasons, multiple alignments of each insertion site were performed using Mafft (http://mafft.cbrc.jp/alignment/software/) and TSDs were identified (Supporting Information Table S1).

HomoeoSNP identification

Homoeologous (A, B and D) genes and their parental orthologs (diploid A, B and D and tetraploid A and B) were aligned (eight genes in total) using Mafft with default parameters and homoeoSNPs were automatically detected using a custom PERL script. Briefly, for each position of the alignment, bases are scored to classify shared homoeoSNPs into three different classes: A/B, A/D and B/D. For each triplet, the total number of homoeoSNPs belonging to each class was calculated and a statistical pairwise binomial test was performed in order to define the homoeology

or subgenome proximity (i.e. relatedness) of the considered triplets. Only triplets with *P*-values < 0.05 were considered for further analysis (see Table S2) and associated to a unique subgenome proximity or relatedness class (A/B or A/D or B/D).

Dating of speciation events

Using the maximum likelihood method in the reference PAML package (Yang, 2007) Ks (synonymous substitution rate) calculation for orthologs/homoeologs between *T. urartu* and *T. aestivum* A subgenome, between *A. speltoides* and *T. aestivum* B subgenome, and between *A. tauschii* and *T. aestivum* D subgenome was performed. The average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site yr^{-1} was used to calibrate the ages of ortholog/homoeolog divergences and then speciation event dates were estimated according to the identification of peaks in Ks distributions. The time (T) of divergence was finally estimated using the formula $T = Ks/2r$.

Results

Syntenome construction to unveil wheat evolution

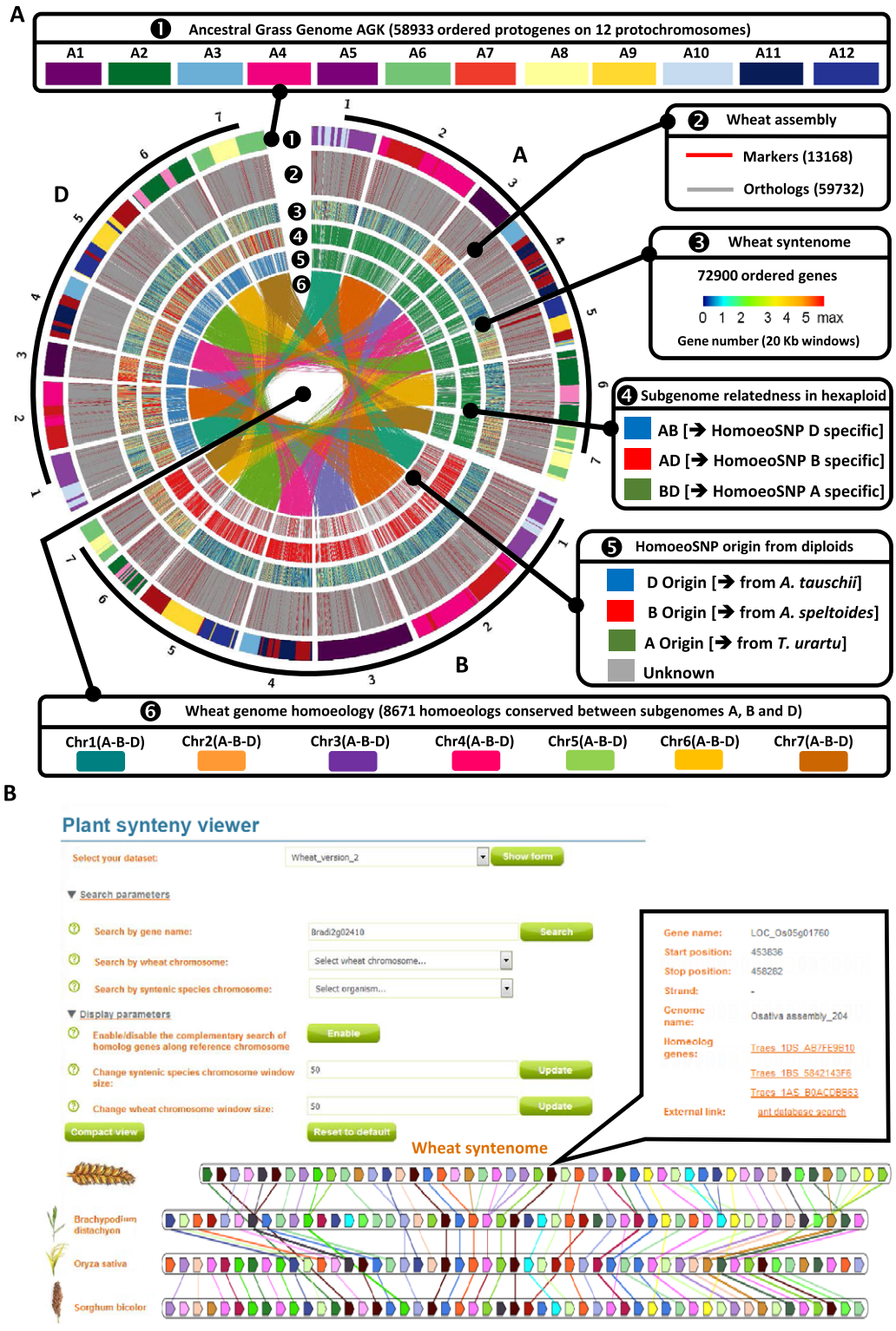
The syntenome is constructed using a synteny-driven approach to order genes/scaffolds on the chromosomes of a species that lacks a physical map. The strategy consists of aligning the ancestral genome (made up of conserved gene adjencies retained in modern species), reconstructed from the lineage of interest (grasses in the current study), to the genetic map of the species of interest (wheat in the current study). The genetic map is then enriched in syntenic (ancestral) genes intercalated between molecular markers, that is, the syntenome (Salse, 2013). From the latest version of the hexaploid wheat genome survey sequence (IWGSC, 2014), consisting of 99 386 gene models (10.2 Mb with 10.8 million scaffolds; Borrill *et al.*, 2015), we produced the most accurate wheat syntenic (also termed 'computed'; Pont *et al.*, 2011, 2013) gene order. Grasses have been proposed to derive from an $n = 7$ ancestor that has been duplicated to reach an $n = 14$ intermediate followed by two chromosomal rearrangements to reach an $n = 12$ ancestor of all modern grasses (Salse, 2016). The $n = 12$ ancestral genome (AGK) consists of 58 933 protogenes (including 17 340 genes conserved between grasses and 41 593 lineage-specific genes), inferred from the comparison of rice, sorghum and *Brachypodium* genomes (Murat *et al.*, 2014; cf. Fig. 1a, circle 1). A total of 13 168 protogenes matched to genetic markers from the most accurate wheat genetic map (Wang *et al.*, 2014) involving 40 267 markers that allowed us to intercalate 59 732 wheat syntenic genes between 13 168 conserved markers (Fig. 1a, circle 2). Overall, based on the chromosome-to-chromosome synteny relationships established between the 21 bread wheat chromosomes and the rice, sorghum and *Brachypodium* genomes, it was then possible to produce the wheat syntenome consisting finally of 72 900 (73.4% of the 99 386 gene models) ordered genes on the 21 chromosomes (Fig. 1a, circle 3), which probably represents the most accurate wheat reference genome available until complete pseudomolecules are publicly released for the 21 chromosomes. We have made this wheat syntenome available through a public web interface named PlantSyntenyViewer at http://urgi.versailles.inra.fr/synteny-wheat (Fig. 1b). Alignment of the 72 900 ordered genes from the wheat syntenome allowed us to identify 8671 robust homoeologous gene triplets (i.e. 26 013 A, B and D gene copies; Fig. 1a, center circle), 5157 pairs (involving 10 314 genes), 15 761 singletons and 10 143 groups of genes (involving 47 298 genes) corresponding to two homologous copies or more but not defining strict homoeologous relationships (i.e. not located on A, B or D subgenomes of the same chromosomal group). Such a syntenome, which allows navigation between grass genomes, can be considered an applied tool for refining structural and functional annotation of wheat orthologous genes, further improving wheat genome sequence assembly, and accelerating identification of candidate genes or markers driving key agronomic traits in wheat (Salse, 2013; Valluru *et al.*, 2014).

TE insertional dynamics during wheat evolution

Following the proposed hybrid origin of the D subgenome (Marcussen *et al.*, 2014; Sandve *et al.*, 2015), TEs shared (located at orthologous positions) between A and B homoeologs (i.e. inserted in their common ancestors) should be observed in the D homoeologous counterpart. Among the 8671 homoeologous gene triplets, 188 exhibit shared TE (class II miniature inverted repeat transposable elements (MITEs) associated with terminal inverted repeats (TIRs)) insertions (Fig. 2a; Table S1). We established that 40, 79 and 69 insertions are shared between, respectively, the A/B, A/D and B/D subgenomes. Precise investigation of the TSD, proof of TE insertion event and then unambiguously rejecting TE excision, established that 16, 43 and 36 insertions are associated with TSDs and shared between, respectively, the A/B, A/D and B/D subgenomes. These results clearly suggest an average of 19%, 43.5% and 37.5% relatedness between the A/B, A/D and B/D wheat subgenomes, respectively. The A (43.5%) and B (37.5%) genomes are more closely related individually to the D genome than to each other (19%). The absence of at least remnants of TSDs (which should remain in the case of TE excision) at precise orthologous sites in the D copy in the case of shared insertions between A and B homoeologs clearly established that 19% of the subgenome D gene-based TEs cannot be inherited exclusively from either the A or B copies. Overall, we precisely identified 19% of shared homoeolog-based TE insertions between the A and B subgenomes, clear proof of an independent origin of the D subgenome that cannot be explained by a pure homoploid origin deriving from the unique hybridization of the A and B progenitors (Marcussen *et al.*, 2014; Sandve *et al.*, 2015), thus reinforcing the hypothesis of a more complex D subgenome origin (Li *et al.*, 2015a,b).

Homoeolog mutational dynamics during wheat evolution

In addition to the previous insertional dynamics of TEs, accumulation of mutations at the gene level should provide additional insights into the origin of the A, B and D wheat subgenomes. In



order to test the accuracy of using homoeoSNP dynamics as a proxy to investigate the origin of the wheat genome, we initially considered the previous 188 homoeologous gene triplets with shared TE insertions for which 19%, 43.5% and 37.5% relatedness between, respectively, the A/B, A/D and B/D subgenomes have been identified (cf. previous section). For the 188 triplets considered, we found that 15%, 44% and 41% of homoeoSNPs were shared between, respectively, the A/B, A/D and B/D subgenomes, a similar rate to that observed for the insertional TE (MITE) fingerprints. The same subgenome affinity was observed when considering the entire set of 8671 homoeologous triplets from the hexaploid bread wheat genome as well as when

Fig. 1 Wheat syntenome. (a) Circle 1, illustration of the synteny between the $n = 12$ ancestral grass karyotype (AGK) (color code for A1–A12) and the 21 bread wheat chromosomes (1–21). Circle 2, illustration of the wheat genes ordered on the 21 chromosomes based on molecular markers (red connecting lines) and synteny with AGK (gray connecting lines). Circle 3, heat map illustration of the gene density (color codes indicate the number of genes within 20-kbp physical windows) on the 21 chromosomes. Circle 4, illustration of subgenome relatedness from the shared mutations (homoeoSNPs) in hexaploid bread wheat (6x) between 8671 homoeologs shown using color codes (blue, red and green for AB, AD and BD relatedness, respectively, or D, B and A specific mutations, respectively). Circle 5, illustration of the gene origin from the homoeoSNP mining strategy for 3121 genes with sequences in diploid (2x), tetraploid (4x) and hexaploid (6x) wheat shown using color codes (blue, red, green and gray for, respectively, D, B, A and unknown origin or, respectively, homoeoSNPs inherited from *Aegilops tauschii*, *Aegilops speltoides*, *Triticum urartu* and none of the three progenitors). Center (Circle 6), illustration of the retained homoeologous triplets (A, B and D copies) on the 21 chromosomes. (b) Screen capture of the PlantSyntenyViewer web tool (http://urgi.versailles.inra.fr/synteny-wheat) showing the synteny between wheat, *Brachypodium*, rice and sorghum and delivering the access to the wheat syntenome consisting of 72 900 genes ordered on the 21 chromosomes.

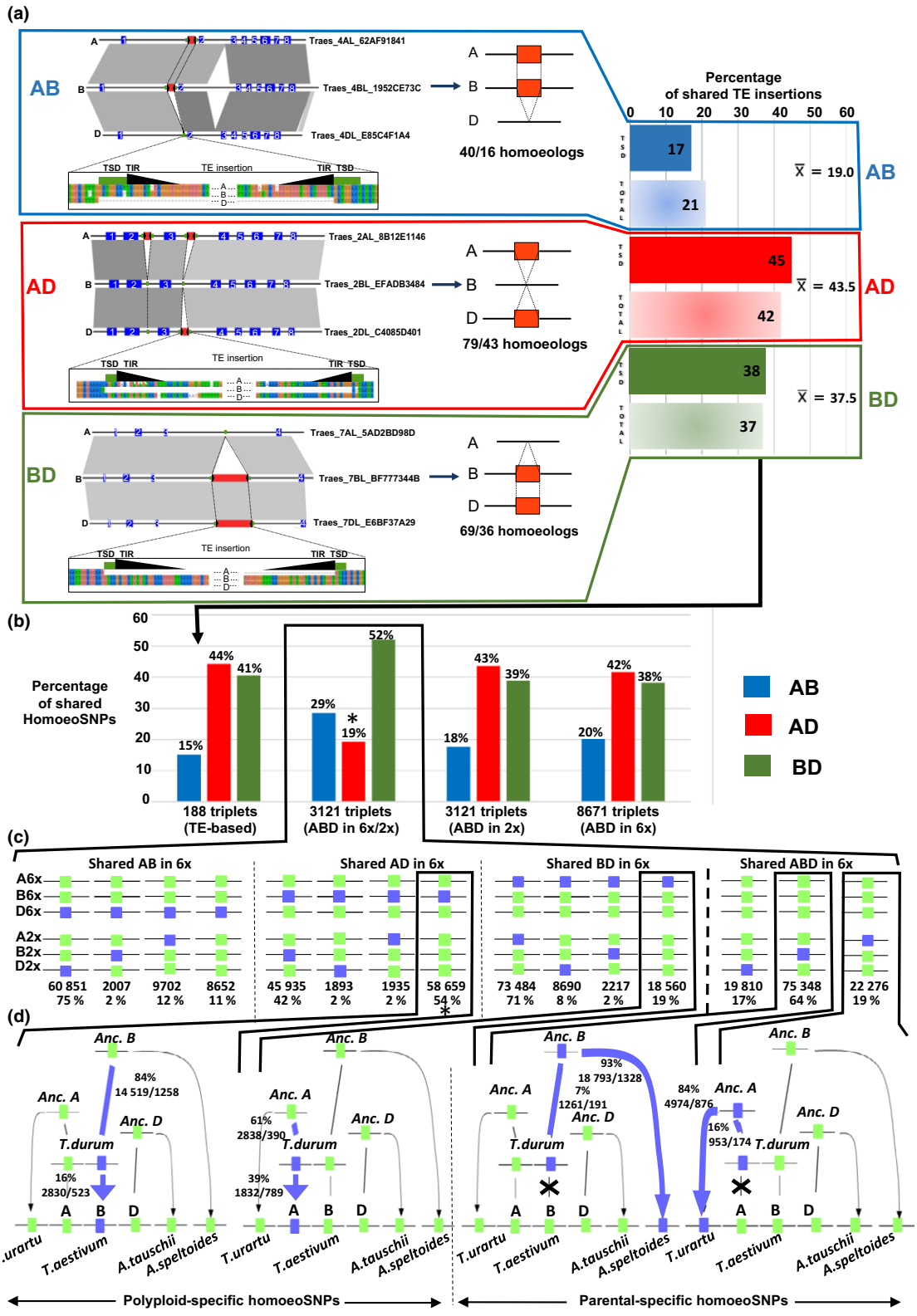
considering the 3121 orthologous genes identified between the diploid (*T. urartu*, *A. speltoides* and *A. tauschii*) progenitors (cf. Figs 2b, 1a, circle 4; Table S2). However, when the 3121 sequence clusters of A, B and D homoeologs from the hexaploid (termed 6x) genome were compared with the orthologous genes in the three considered progenitors (termed 2x), a clear depletion in A/D sequence affinity was observed (Fig. 2b, black asterisks).

While 75% and 71% of homoeoSNPs observed, respectively, in the A and D subgenomes of hexaploid bread wheat were inherited from their founder progenitors (respectively *T. urartu* and *A. tauschii*), only 42% of homoeoSNPs located in the B subgenome were derived from *A. speltoides* (Figs 2c, 1a, circle 5). In order to orientate the changes in homoeoSNP accumulation during evolution, we compared groups of orthologous genes associated with three copies in the hexaploid (6x), two copies in the tetraploid (4x) and one copy in the three diploid progenitors (2x) (Fig. 2d). Defining such clusters of eight orthologous genes (three from the hexaploid, two from the tetraploid and one from each of the three diploids) allows us to assess the transmission of mutations during evolution from the diploid to the tetraploid and finally to the hexaploid, ultimately defining homoeoSNPs between the A, B and D subgenomes. Taking into account that the exact founder diploid individual(s) will never be known and that the progenitors and their resultant polyploids (4x and 6x) may have evolved differentially through differences in mutation rates, genetic drift, genetic admixture or may even have experienced distinct rounds of domestication, perfect homoeoSNP inheritance between 2x, 4x and 6x wheats is not expected. Eighty-four per cent of homoeoSNPs observed in the B subgenome in the hexaploid (6x), but not inherited (absent) from *A. speltoides* (2x), were identified in the B subgenome of the tetraploid (4x), thus making the remaining homoeoSNPs (16%) specific from the B subgenome in the hexaploid. In comparison, 61% of homoeoSNPs observed in the A subgenome in the hexaploid (6x), but not inherited from *T. urartu* (2x), were identified in the A subgenome of the tetraploid (4x), thus making 39% of such homoeoSNPs specific from the A subgenome in the hexaploid. This suggests a more ancient origin of the B progenitor (84% of B homoeoSNPs acquired between 2x and 4x) compared with the A progenitor (61% of A homoeoSNPs acquired between 2x and 4x), or, more precisely, a more ancient speciation between *A. speltoides* (2x)/B subgenome (6x and 4x) compared with *T. urartu* (2x)/A subgenome (6x and 4x). This conclusion is supported by the mutations identified in the progenitors (*T. urartu* and *A. speltoides*) not transmitted to hexaploid bread wheat. Ninety-three per cent of the mutations identified in *A. speltoides* were not transmitted to the tetraploid (4x), and thus consist of lineage-specific mutations accumulated in *A. speltoides* since its divergence from the tetraploid (AB) progenitor. The remaining mutations (7%) consist of homoeoSNPs shared between *A. speltoides* and the tetraploid subgenome B but not transmitted to the hexaploid subgenome B as a result probably of random (and few) substitutions, deletions or alternatively gene conversions between homoeologs. In comparison to 84% *T. urartu* lineage-specific mutations identified (i.e. mutations from *T. urartu* not transmitted to the tetraploid), the number of *A. speltoides* mutations that were either transmitted to the tetraploid/hexaploid wheat (i.e. ancestral) or lineage-specific (i.e. not transmitted) ones, supports a more ancient origin of the B progenitor compared with the A and D progenitors.

Wheat evolutionary model

Based on the evolutionary dynamics at the TE and mutation levels, we propose a novel model of hexaploid bread wheat origin (Fig. 3). In this scenario (from top to bottom), from the three ancestral progenitors (termed AncA, AncS and AncD), whereas the evolution of the A subgenome from hexaploid bread wheat appears quite simple, the evolution of the other two subgenomes is more complex than initially reported. The A subgenome in tetraploid/hexaploid wheat derived from AncA and diverged from the modern *T. urartu* and *T. monococcum* AncA representatives, respectively, 0.23–0.46 Ma. The B subgenome in tetraploid/hexaploid wheat derived from a more ancient AncS progenitor and diverged from the modern *A. speltoides* AncS representative 1.15–2.7 Ma. Regarding the D subgenome in tetraploid/hexaploid wheat, it derived from a complex history (multiple rounds) of hybridization between AncA and AncS but also with another specific progenitor (termed AncD and accounting for at least 19% of the origin of the modern D subgenome) that diverged from the modern *A. tauschii* representative 0.07–0.3 Ma (Fig. 3).

A particular pattern of mutation accumulation has thus been observed in the B subgenome, presented previously as proof of a more ancient origin of the B progenitor, or more precisely an ancient speciation between the B subgenome in the tetraploid/hexaploid and *A. speltoides* (considered as a modern representative of AncB). However, another explanation has been proposed introducing a possible polyphyletic origin of AncB resulting from an introgression of several parental *Aegilops* species from the



Sitopsis section (termed S and including *Aegilops bicornis*, C^b; *Aegilops searsi*, S^s; *Aegilops longissimi*, S^l; *Aegilops sharonensis*, S^h; *Aegilops speltoides*, S) that need to be identified, if they are not extinct. However, the two previous scenarios for the origin of the B subgenome (i.e. ancient and/or polyphyletic origins) rely on the assumption of a constant and similar evolutionary rate in the subgenomes after polyploidization so that the observed modern mutations were inherited from the progenitor(s) from the *Sitopsis* section, and/or gained/lost at a constant rate in the subgenomes in the course of evolution. The current data allowed us to reveal

Fig. 2 Transposable element (TE) and homoeoSNP evolutionary dynamics. (a) (left) Illustration of the identified TEs shared between A and B (upper), A and D (middle) and B and D (lower) homoeologs (exons in blue with numbers) defining sequence conservation (gray blocks) breaks (illuminated by the sequence alignment) defining target site duplication (TSD) and terminal inverted repeat (TIR) elements. (right) Illustration of the observed percentage (and associated mean value $\bar{X}$) of shared TEs between A and B (upper), A and D (middle) and B and D (lower) homoeologs, with dark-colored bars for the total shared TE insertions and light-colored bars for shared TE insertions with TSD (green boxes)/TIR (black boxes). (b) Percentage of shared homoeoSNPs between A and B (blue bars), A and D (red bars) and B and D (green bars) for 188 gene triplets from the hexaploid bread wheat (associated with shared TEs), 3121 gene triplets from the hexaploid bread wheat (associated with orthologs in the diploid progenitors) and 8671 gene triplets from the hexaploid bread wheat (total homoeolog repertoire). The black asterisks indicate depletion in A/D sequence affinity. (c) Illustration of the homoeoSNP conservation between the hexaploid bread wheat subgenomes (termed A6x, B6x and D6x) and the diploid progenitors *Triticum urartu* (A2x), *Aegilops speltoides* (B2x) and *Aegilops tauschii* (D2x). The number of homoeoSNPs and associated percentages are shown at the bottom of each illustrated case. (d) Illustration of the evolutionary models supporting the observed 'polyploid-specific' homoeoSNPs (left, with the B subgenome on the left and the A subgenome on the right) and 'parental-specific' homoeoSNPs (right, with the B subgenome on the left and the A subgenome on the right). The blue lines indicate the different cases of homoeoSNP transmissions (with percentage followed by the number of homoeoSNPs and genes involved) from the parents (2x) to the tetraploid (4x) and hexaploid (6x) genomes. Crosses indicate deletions. Anc., ancestor.

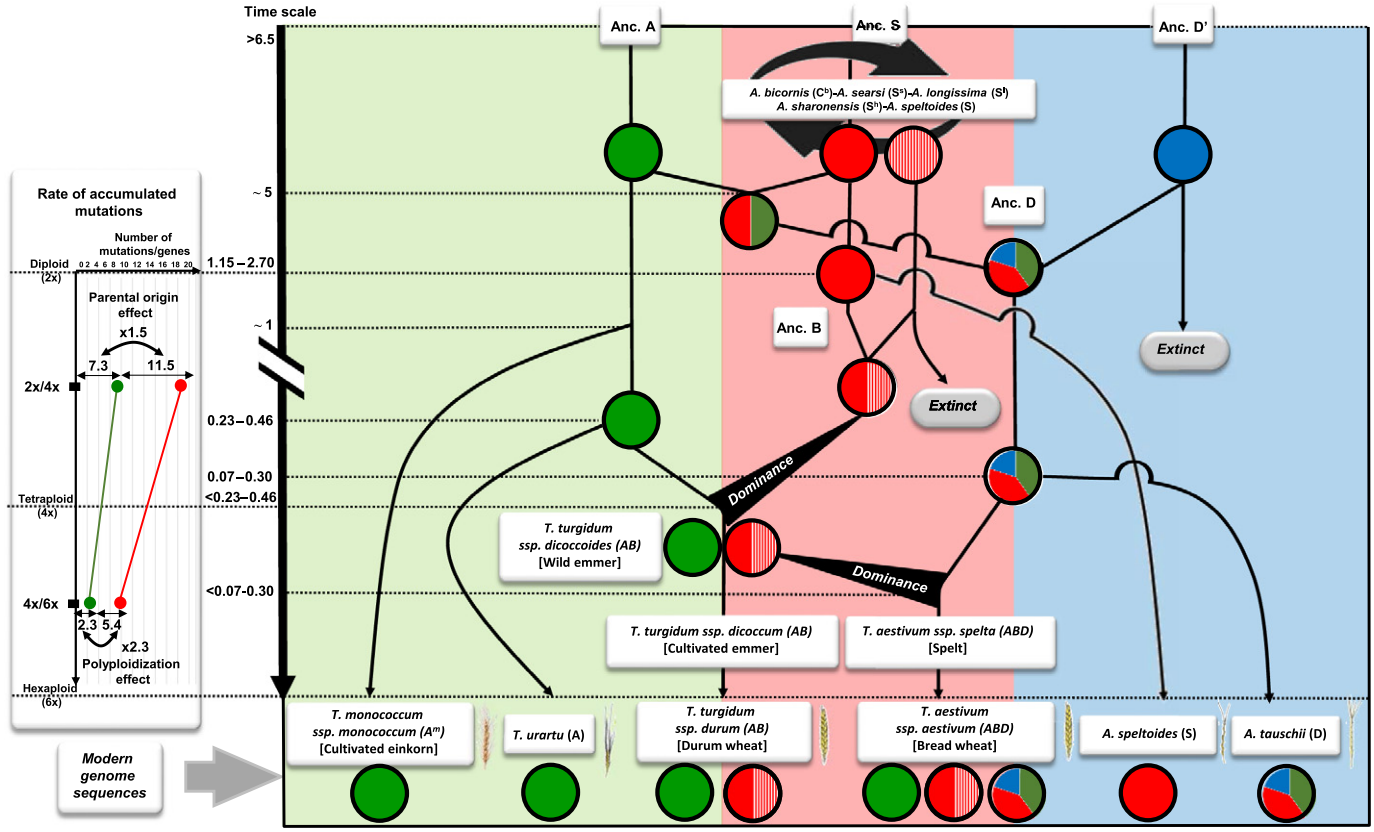


Fig. 3 Wheat evolutionary model. Illustration of paleohistory of hexaploid bread wheat from ancestor genome A (Anc. A; green circle), ancestor genome B (Anc. B; red circle; derived from the hybridization of *Siptosis* species, indicated by a circular black arrow) and ancestor genome D (Anc. D; blue circle) with the time scale being in million years (left). Subgenomes are illustrated with circles so that hybridization events are highlighted with mixed color within circles. Subgenome dominance following polyploidization is illustrated with large black arrows indicating the accelerated accumulation of mutations. Modern sequenced species are illustrated at the bottom of the figure. The rates of mutation accumulation observed in the A and B subgenomes for the transition between the parents and the tetraploid (transition 2x to 4x) and between the tetraploid and the hexaploid (transition 4x to 6x) are shown on the left.

precisely the rate of homoeoSNP accumulation in the modern bread wheat subgenomes inherited from the parents and/or from the polyploidization events. Taking into account the 'polyploid-specific homoeoSNPs' (Fig. 2, left panel), that is, homoeoSNPs identified in the A subgenome in the hexaploid and absent from *T. urartu*, 7.3 homoeoSNPs/genes (i.e. 2838 homoeoSNPs in 390 genes with an average size of 4.04 kbp per gene) originated from the transition between the diploid and the tetraploid (termed 2x to 4x) and 2.3 homoeoSNPs/genes (i.e. 1832 homoeoSNPs in 789 genes with an average size of 3.75 kbp per gene) from the transition between the tetraploid and the hexaploid (termed 4x to 6x). In the same manner, for the B subgenome, that is, homoeoSNPs observed in the B subgenome in the hexaploid and absent from *A. speltoides*, 11.5 homoeoSNPs/genes (i.e. 14 519 homoeoSNPs in 1258 genes with an average size of 3.66 kbp per gene) originated from the transition from 2x to 4x and 5.4 homoeoSNPs/genes (i.e. 2830 homoeoSNPs in 523 genes with an average size of 3.98 kbp per

gene) from the transition between 4x and 6x. Comparing now the accumulation rate of homoeoSNPs per genes (with genes of similar size as a clear proof of homoeoSNPs density/rate consistency) for the two considered transitions (2x to 4x and 4x to 6x), we observed an accelerated rate of homoeoSNP accumulation for the B subgenome compared with the A subgenome between 2x and 4x (rate of 1.5x = 11.5/7.3) and between 4x and 6x (rate of 2.3x = 5.4/2.3). Whereas the observed 1.5x higher accumulation rate of homoeoSNPs in the B subgenome (compared with A) during the 2x-to-4x transition was explained previously by the ancient mono- or polyphyletic origin of the B progenitor, the observed net 2.3x increase in homoeoSNP accumulation in the B subgenome during the 4x-to-6x transition is consistent with the recently proposed contrasting plasticity (i.e. dominance or partitioning) of the subgenomes following polyploidization in wheat (Pont *et al.*, 2013) and more generally in plants (Murat *et al.*, 2014, 2015a,b).

Discussion

An investigation of the evolutionary dynamics of TEs and mutations in wheat allowed us to propose a scenario in which the A subgenome derived from an ancestral genome closely related to the modern *T. urartu*, with 71% of mutations detected in the ancestor (AncA) transmitted to the modern subgenome; the B genome derived from an ancient mono- or polyphyletic progenitor and experienced accelerated plasticity following polyploidization, which together explained why only 42% of mutations identified in the modern subgenome were inherited from the ancestral genome (AncB) closely related to the modern *A. speltoides*; and the D subgenome derived from a complex hybridization pattern of the three A, B and D progenitors accounting, respectively, for on average 38%, 43% and 19% of the modern D subgenome in hexaploid bread wheat, based on TE (from 188 triplets) and mutation (from 8671 triplets) insertional dynamics. The current model first reconciles data from previous studies addressing the origin of subgenome D, as our results support the conclusions of two recent studies suggesting that the D subgenome has a homoploid origin (Marcussen *et al.*, 2014; Sandve *et al.*, 2015). The hybridization of the A and B ancestors as proposed in those studies does not entirely explain the origin of the modern D subgenome of hexaploid bread wheat that also derived from a specific D progenitor independent from A and B ancestors (Li *et al.*, 2015a,b).

Regarding the controversial B subgenome paleohistory, it becomes clear that the proposal of an ancestral (more ancient than the A and D progenitors) mono- or polyphyletic B subgenome origin cannot explain entirely the observed accumulation of mutations during evolution in shaping the modern bread wheat B subgenome. By contrast, we propose an alternative scenario where the increased divergence of the B subgenome in the hexaploid wheat compared to *A. speltoides* at the sequence (homoeoSNPs) level is the consequence of a differential evolutionary plasticity of the B subgenome compared with the A and D subgenomes in response to polyploidization events. Polyploidization has been shown to be followed by a subgenome dominance phenomenon with contrasting plasticity of the post-duplication blocks leading, at the whole-chromosome or genome level, to dominant (D; retention of duplicated genes; also termed least fractionated (LF)) and sensitive (S; loss of duplicated genes; also termed most fractionated (MF)) compartments (Salse, 2016). Such subgenome dominance following polyploidization has been reported in Arabidopsis (Thomas *et al.*, 2006), maize (*Zea mays*) (Woodhouse *et al.*, 2010; Schnable *et al.*, 2012a,b), and *Brassica* (Cheng *et al.*, 2012). In the Brassicaceae, such subgenome dominance has been proposed following the *Brassica rapa* hexaploidization between the three post-polyploidy compartments termed the LF, medium fractionated (MF1) and most fractionated (MF2) blocks. The contrasting plasticity between the MF and LF compartments in *B. rapa* has been associated with bias in (1) gene retention and with genes retained in pairs or triplets enriched in functional categories such as transcriptional regulation, ribosomes, response to abiotic or biotic stimuli, response to hormonal stimuli, cell organization and transporter functions; (2) gene expression, with genes located in the LF subgenome proposed to be dominantly expressed over those located in the two proposed fractionated subgenomes (MF1 and MF2); and (3) single nucleotide polymorphism (SNP) at the population level, with genes located in LF showing fewer nonsynonymous or frameshift mutations than genes in MF fractions (Edger & Pires, 2009; Cheng *et al.*, 2012, 2013, 2014; Fang *et al.*, 2012).

Here, our findings also support a recent evolutionary scenario introducing the concept of subgenome dominance between the A, B and D subgenomes of wheat following polyploidizations (Pont *et al.*, 2013), where the modern bread wheat genome has been shaped by a first neotetraploidization event (< 0.5 Ma) leading to subgenome dominance where the A subgenome was dominant and the B subgenome sensitive (i.e. prone to the observed mutation accumulation). The second neohexaploidization event (< 0.3 Ma) led potentially to a supra-dominance where the tetraploid became sensitive (subgenomes A and B) and the D subgenome dominant (i.e. pivotal).

The current study offers new insights into the origin of modern bread wheat. The historical evolution of the bread wheat genome appears to be far more complex than initially suggested. More complete genome sequences from diploid, tetraploid and hexaploid wheat will offer the opportunity, in the long term, to improve upon the currently proposed evolutionary scenario to explain how the modern bread wheat genome has been shaped from its diploid progenitors.

Acknowledgements

This work has been supported by grants from INRA ('Génétique et Amélioration des Plantes, ref: 'Appel d'Offre Front de Science' projet TransWHEAT), the Agence Nationale de la Recherche (program ANR Blanc-PAGE, ref: ANR-2011-BSV6-00801), the Agreenskills program ('TransGRAIN', session 2014, ID: 459) and the 'Région Auvergne, Allocation de recherche Territoire, Agriculture, Alimentation, Nutrition et Santé Humaine' (contract no. 23000720).

Author contributions

M.E.B., F.M., M.V., M.M. and C.P. participated in the data analysis as well as in preparation of the manuscript; R.F., L.B., M.A. and H.Q. developed and managed the PlantSyntenyViewer web tool; C.P. and J.S. managed the research project; J.S. wrote the manuscript.

References

Borrill P, Adamski N, Uauy C. 2015. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist* **208**: 1008–1022.

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X. 2013. Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. *Plant Cell* **25**: 1541–1554.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE* **7**: e36442.

Cheng F, Wu J, Wang X. 2014. Genome triplication drove the diversification of *Brassica* plants. *Horticulture Research* **1**: 14024.

Devos KM, Dubcovsky J, Dvorák J, Chinoy CN, Gale MD. 1995. Structural evolution of wheat chromosomes 4A, 5A and 7B and its impact on recombination. *Theoretical and Applied Genetics* **91**: 282–288.

Dvorak J, Akhunov ED. 2005. Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance. *Genetics* **171**: 323–332.

Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Molecular Biology and Evolution* **23**: 1386–1396.

Dvorak J, Zhang HB. 1990. Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. *Proceedings of the National Academy of Sciences, USA* **87**: 9640–9644.

Dvorak J, Zhang HB, Kota RS, Lassner M. 1989. Organization and evolution of the 5S ribosomal RNA gene family in wheat and related species. *Genome* **32**: 1003–1016.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* **17**: 699–717.

Fang L, Cheng F, Wu J, Wang X. 2012. The impact of genome triplication on tandem gene evolution in *Brassica rapa*. *Frontiers in Plant Science* **3**: 261.

Feldman M. 1966a. Identification of unpaired chromosomes in F_1 hybrids involving *Triticum aestivum* and *T. timopheevii*. *Canadian Journal of Genetics and Cytology* **8**: 144–151.

Feldman M. 1966b. The mechanism regulating pairing in *Triticum timopheevii*. *Wheat Information Service* **21**: 1–2.

Feldman M, Lupton FGH, Miller TE. 1995. Wheats. In: Smartt J, Simmonds NW, eds. *Evolution of crop plants, 2nd edn*. Harlow, UK: Longman Scientific & Technical, 184–192.

Gill BS, Chen PD. 1987. Role of cytoplasm specific introgression in the evolution of the polyploid wheats. *Proceedings of the National Academy of Sciences, USA* **84**: 6800–6804.

Huang S, Sirikhachornkit A, Su X, Faris J, Gill B, Haselkorn R, Gornicki P. 2002. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences, USA* **99**: 8133–8138.

Hutchinson J, Miller TE, Jahier J, Shepherd KW. 1982. Comparison of the chromosomes of *Triticum timopheevii* with related wheats using the techniques of C-banding and in situ hybridization. *Theoretical and Applied Genetics* **64**: 31–40.

International Brachypodium Initiative (IBI). 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763–768.

International Rice Genome Sequencing Project (IRGSP). 2005. The map-based sequence of the rice genome. *Nature* **436**: 793–800.

International Wheat Genome Sequencing Consortium (IWGSC). 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**: 1251788.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X *et al.* 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**: 91–95.

Jiang J, Gill BS. 1994. Different species-specific chromosome translocations in *Triticum timopheevii* and *T. turgidum* support the diphyletic origin of polyploid wheats. *Chromosome Research* **2**: 59–64.

Kilian B, Ozkan H, Deusch O, Effgen S, Brandolini A, Kohl J, Martin W, Salamini F. 2007. Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. *Molecular Biology and Evolution* **24**: 217–227.

Li LF, Liu B, Olsen KM, Wendel JF. 2015a. A re-evaluation of the homoploid hybrid origin of *Aegilops tauschii*, the donor of the wheat D-subgenome. *New Phytologist* **208**: 4–8.

Li LF, Liu B, Olsen KM, Wendel JF. 2015b. Multiple rounds of ancient and recent hybridizations have occurred within the *Aegilops–Triticum* complex. *New Phytologist* **208**: 11–12.

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y *et al.* 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**: 87–90.

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S *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. *Proceedings of the National Academy of Sciences, USA* **110**: 7940–7945.

Maestra B, Naranjo T. 1999. Structural chromosome differentiation between *Triticum timopheevii* and *T. turgidum* and *T. aestivum*. *Theoretical and Applied Genetics* **98**: 744–750.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, International Wheat Genome Sequencing Consortium, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF *et al.* 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**: 1250092.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Roest Crollius H, Salse J. 2015a. Understanding *Brassicaceae* evolution through ancestral genome reconstruction. *Genome Biology* **16**: 262.

Murat F, Zhang R, Guizard S, Flores R, Armero A, Pont C, Steinbach D, Quesneville H, Cooke R, Salse J. 2014. Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biology and Evolution* **6**: 12–33.

Murat F, Zhang R, Guizard S, Gavranović H, Flores R, Steinbach D, Quesneville H, Tannier E, Salse J. 2015b. Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. *Genome Biology and Evolution* **7**: 735–749.

Naranjo T. 1990. Chromosome structure of durum wheat. *Theoretical and Applied Genetics* **79**: 397–400.

Naranjo T, Roca A, Goicoechea PG, Giráldez R. 1987. Arm homoeology of wheat and rye chromosomes. *Genome* **29**: 873–882.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.

Pham SK, Pevzner PA. 2010. DRIMM-synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics* **26**: 2509–2516.

Pont C, Murat F, Confolent C, Balzergue S, Salse J. 2011. RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.). *Genome Biology* **12**: R119.

Pont C, Murat F, Guizard S, Flores R, Foucrier S, Bidet Y, Quraishi UM, Alaux M, Doležel J, Fahima T *et al.* 2013. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant Journal* **76**: 1030–1044.

Salina EA, Lim KY, Badaeva ED, Shcherban AB, Adonina IG, Amosova AV, Samatadze TE, Vatolina TY, Zoshchuk SA, Leitch AR. 2006. Phylogenetic reconstruction of *Aegilops* section *Sitopsis* and the evolution of tandem repeats in the diploids and derived wheat polyploids. *Genome* **49**: 1023–1035.

Salse J. 2013. Paleogenomics as a guide for traits improvement: volume 1. Managing, sequencing and mining genetic resources. In: Tuberosa R, Graner A, Frison E, eds. *Genomics of plant genetic resources*. Dordrecht, the Netherlands: Springer, 131–172.

Salse J. 2016. Ancestors of modern plant crops. *Current Opinion in Plant Biology* **30**: 134–142.

Salse J, Abrouk M, Murat F, Quraishi UM, Feuillet C. 2009. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Briefings in Bioinformatics* **10**: 619–630.

Salse J, Chagué V, Bolot S, Magdelenat G, Huneau C, Pont C, Belcram H, Couloux A, Gardais S, Evrard A *et al.* 2008. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics* **25**: 555.

Sandve SR, Marcussen T, Mayer K, Jakobsen KS, Heier L, Steuernagel B, Wulff BB, Olsen OA. 2015. Chloroplast phylogeny of *Triticum/Aegilops* species is not incongruent with an ancient homoploid hybrid origin of the ancestor of the bread wheat D-genome. *New Phytologist* **208**: 9–10.

Schnable JC, Freeling M, Lyons E. 2012a. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biology and Evolution* **4**: 265–277.

Schnable JC, Wang X, Pires JC, Freeling M. 2012b. Escape from preferential retention following repeated whole genome duplications in plants. *Frontiers in Plant Science* **3**: 94.

Terachi T, Ogihara Y, Tsunewaki K. 1990. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. 7. Restriction endonuclease analysis of mitochondrial DNA from polyploid wheats and their ancestral species. *Theoretical and Applied Genetics* **80**: 366–373.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**: 934–946.

Valluru R, Reynolds MP, Salse J. 2014. Genetic and molecular bases of yield-associated traits: a translational biology approach between rice and wheat. *Theoretical and Applied Genetics* **127**: 1463–1489.

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L *et al.* 2014. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnology Journal* **12**: 787–796.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biology* **8**: e1000409.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Zohary D, Feldman M. 1962. Hybridization between amphidiploids and the evolution of polyploids in the wheat (*Aegilops-Triticum*) group. *Evolution* **16**: 44–61.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Table S1 TE repertoire

Table S2 HomoeoSNP repertoire

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

ANNEXE 2

Supplementary data de l'article (Chapitre 1)

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) doi: 10.1111/tpj.12366

Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes

Caroline Pont[1,†], Florent Murat[1,†], Sébastien Guizard[1], Raphael Flores[2], Séverine Foucrier[1], Yannick Bidet[3], Umar Masood Quraishi[1], Michael Alaux[2], Jaroslav Doležel[4], Tzion Fahima[5], Hikmet Budak[6], Beat Keller[7], Silvio Salvi[8], Marco Maccaferri[8], Delphine Steinbach[2], Catherine Feuillet[1], Hadi Quesneville[2] and Jérôme Salse[1,*]

SUPPORTING FIGURES

Figure S1: ***Wheat syntenome characterization flow chart***. Schematic representation of the performed analysis in characterizing the wheat syntenome then leading to subgenome dominance investigation. The public grass genomes have been compared to derive an exhaustive set of 17,317 COS (top, brown panel). Computed gene order protocol delivered a wheat syntenome where in average ~2400 genes are ordered per wheat chromosome group (middle purple panel). The strategy #1, in investigating the structural difference between subgenomes A, B and D, was to align the characterized wheat syntenome to the bread wheat genome sequence (Brenchley et al 2012) in delivering wheat homoeolog gene set. The second complementary strategy #2 consisted in sequencing wheat ordered COS genes in order to indentify RTs (reference transcripts delivering homoeologous relationships in Chinese Spring) and associated SNPs/COS density (in sequencing eight hexaploid wheat genotypes from the characterized RTs). Both complementary strategies allowed the fine characterization of structural bias in gene content (*i.e.* COS content), gene diversity (*i.e.* SNPS/COS), gene transposition (*i.e.* non-syntenic genes), recombination profile (cM/marker) and C-banding (using directly the C-band information per BIN from Hutchinson *et al.,* 1982) between wheat paleo- as well as neodominant and sensitive blocks.

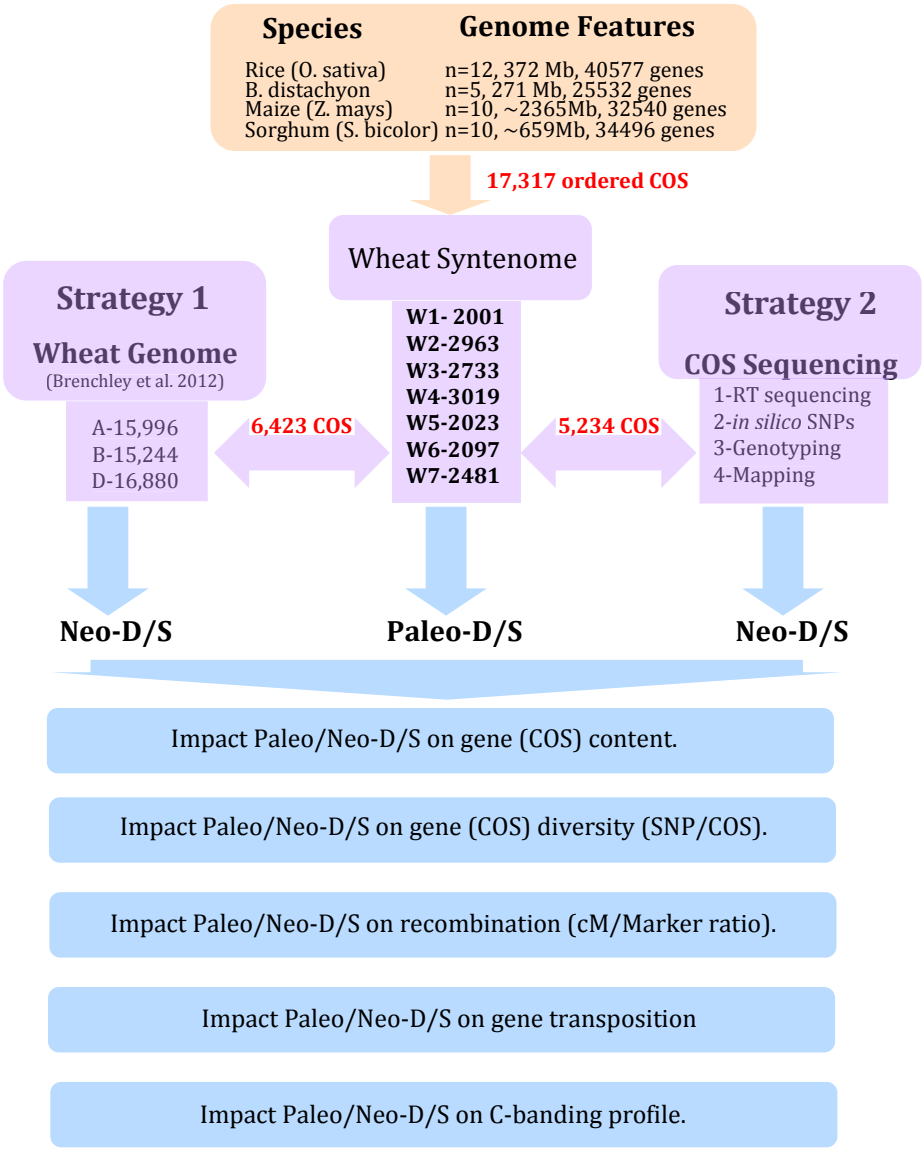
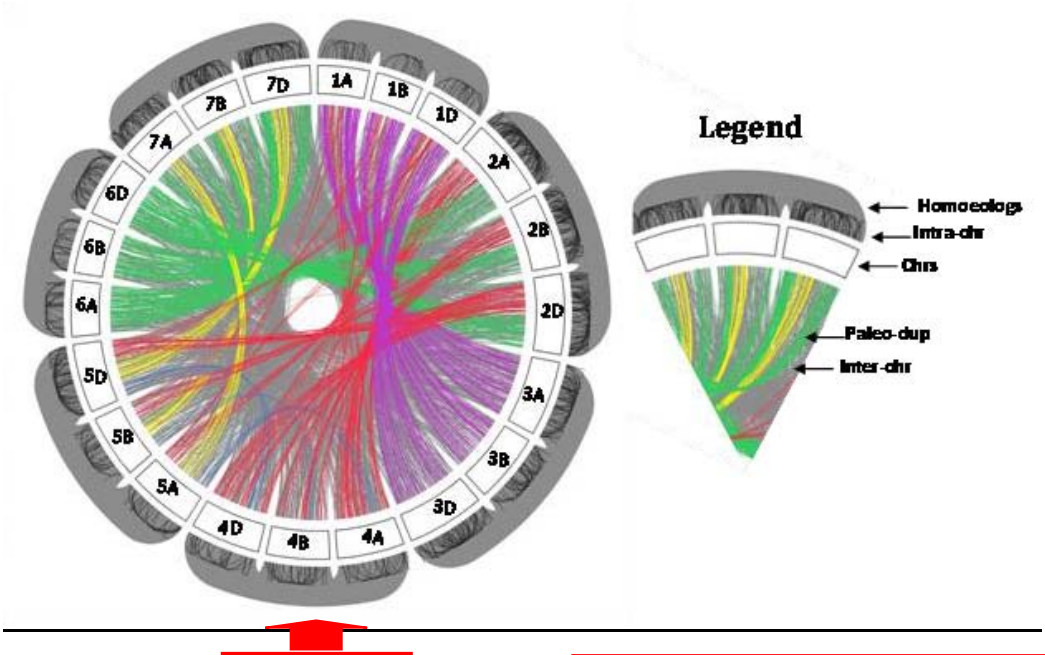


Figure S2: ***Wheat duplicated genes (homoeologs & paralogs) characterization***. **A-** Illustration on the 21 bread wheat chromosomes (from 1A to 7D), shown as a circle, of the homoeologs [duplicated genes between subgenomes A, B and D with 6062 genes located on AB (5063), AD (5557), BD (5252)] as well as intra-chromosomal [duplicated genes within chromosomes], paleochromosomal [ancestral WGD shown in color connecting lines], inter-chromosomal [between chromosomes groups 1 to 7 excluding ancestral WGD paralogs and shown in grey connecting lines] duplicated genes (see legend at the right) associated with 6,423 ordered protogenes (syntenome). **B-** The table delivers the number of paralogous genes classified into intra-chromosomal, paleochromosomal, inter-chromosomal duplications mapped on the A, B and D subgenomes, as illustrated on the wheat genome circle.

A



B

	Brenchley et al.	6423 COS	Paralogs	Intra-chr	Inter-chr	Paleo-dup
A	15996	5827	1296	237	826	233
B	15244	5472	1208	217	794	197
D	16880	6091	1360	257	847	256

Figure S3: ***Grass dominant and sensitive compartments following ancestral WGD***. Grass sub-genome portioning following ancestral WGD is illustrated with rice as the modern grass genome closest representative of the n=12 ancestor intermediate (A1 to A12 illustrated as black bars) deriving from the n=7 ancestor (top). The distribution (bottom) illustrates the number of ancestral retained genes (y-axis) in paralogous blocks observed in the modern rice genome (chromosomes r1 to r12, x-axis). Dominant and sensitive chromosomal blocks are illustrated respectively as blue and red vertical bars.

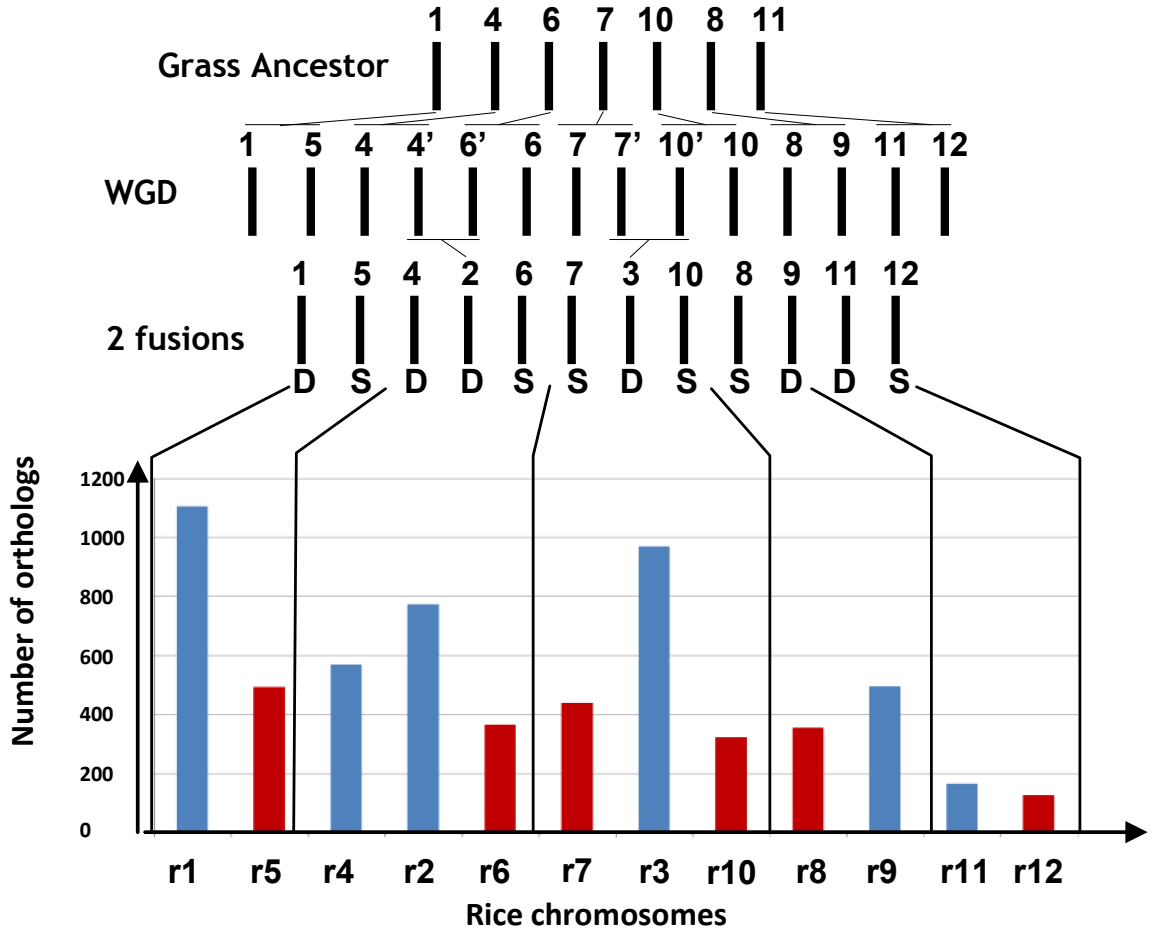


Figure S4: ***Strategy for COS-SNP marker development and exploitation.*** **A-** Schematic representation of the COS sequencing strategy aiming at identifying SNP markers. Step 1 - Synteny based COS identification using COS-*finder* software. Step 2 - Amplicon/Sequence-capture based long reads sequencing in reference genotype. Step 3 - Short reads sequencing in diversity panel and SNP calling. Step 4 - Genotyping. Step 5 - Mapping. The strategy (Figure S3) consisted in sequencing the COS set using primers defined in the previous section on a reference genotype to identify homoeoSNPs (Step 1). Amplicons sequencing (454 Roche, for longer reference reads) and sequence clustering using parameters described in the experimental procedure section allow the identification of reference transcripts RTs (putatively homoeologs) (Step 2). Finally, sequencing (short reads such as solexa) of additional genotypes (either based on amplicon or sequence capture strategies) and reads mapping using tools described in the experimental procedure section allow the identification of *in silico* SNPs for all the homoeologs detected in the previous step (Step 3). COS-SNP have been then genotyped in several ways as described in the experimental procedure section (Step 4, Figure S4) and finally mapped based on appropriate recombinant population (Step 5). **B-** Schematic representation of the different low, medium and high throughput genotyping technologies that can be applied with the provided COS-SNP markers such as size detection (on capillary sequencer ABI 3730), Illumina veracode detection, High Resolution Melting (HRM) analysis, LNA detection, KASpar detection, Single Strand Conformation Polymorphism (SSCP) detection (see experimental procedure section). A single COS (COS-5368) has been sequenced between *cv* Chinese Spring (CS) and *cv* Renan (Re) for the three homoeologs delivering the A/G SNP on the D homoeoallele that has been used to illustrate the different SNP genotyping approaches.

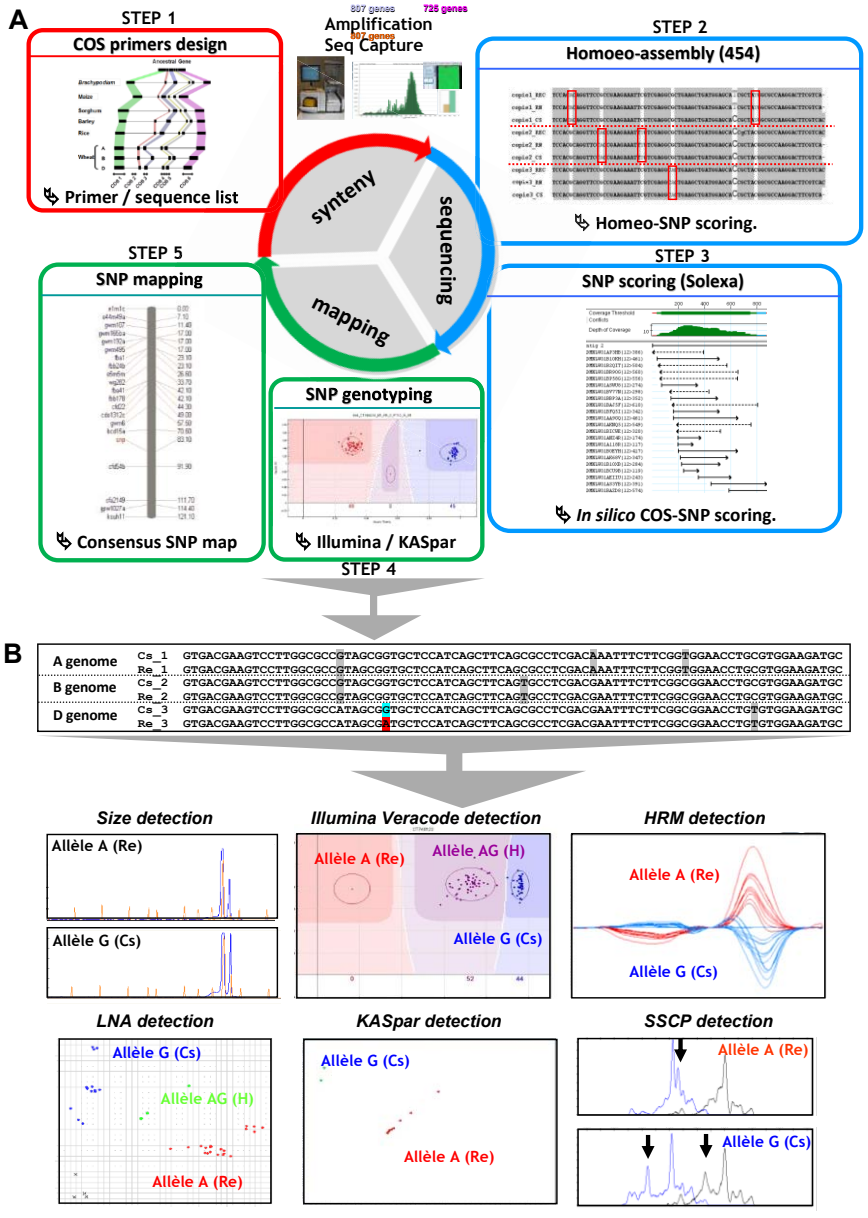


Figure S5: ***COS-SNP transferability in sorghum, maize, oat, rice, Triticale, wheat, wheat, rye, barley, ray grass (illustrated with COS-5598).*** **A -** Agarose gel illustrates polymorphism in intron size. Monocot species are shown at the left of the gel. **B -** Fragments visualisation on capillary sequencer ABI 3730XL allows to visualize short size polymorphism. The X-axis represents the fragment sizes, the Y axis represents the peak (*i.e.* fluorophore) intensities. **C -** High Resolution Melting (HRM) profile obtained by PCR amplifications and validation through sequence analysis (top and bottom sequence haplotypes) for a single COS marker in six ray grass species (mentioned on the phylogenetic tree) allow to confirm SNPs (*i.e.* differences between red and blue curves).

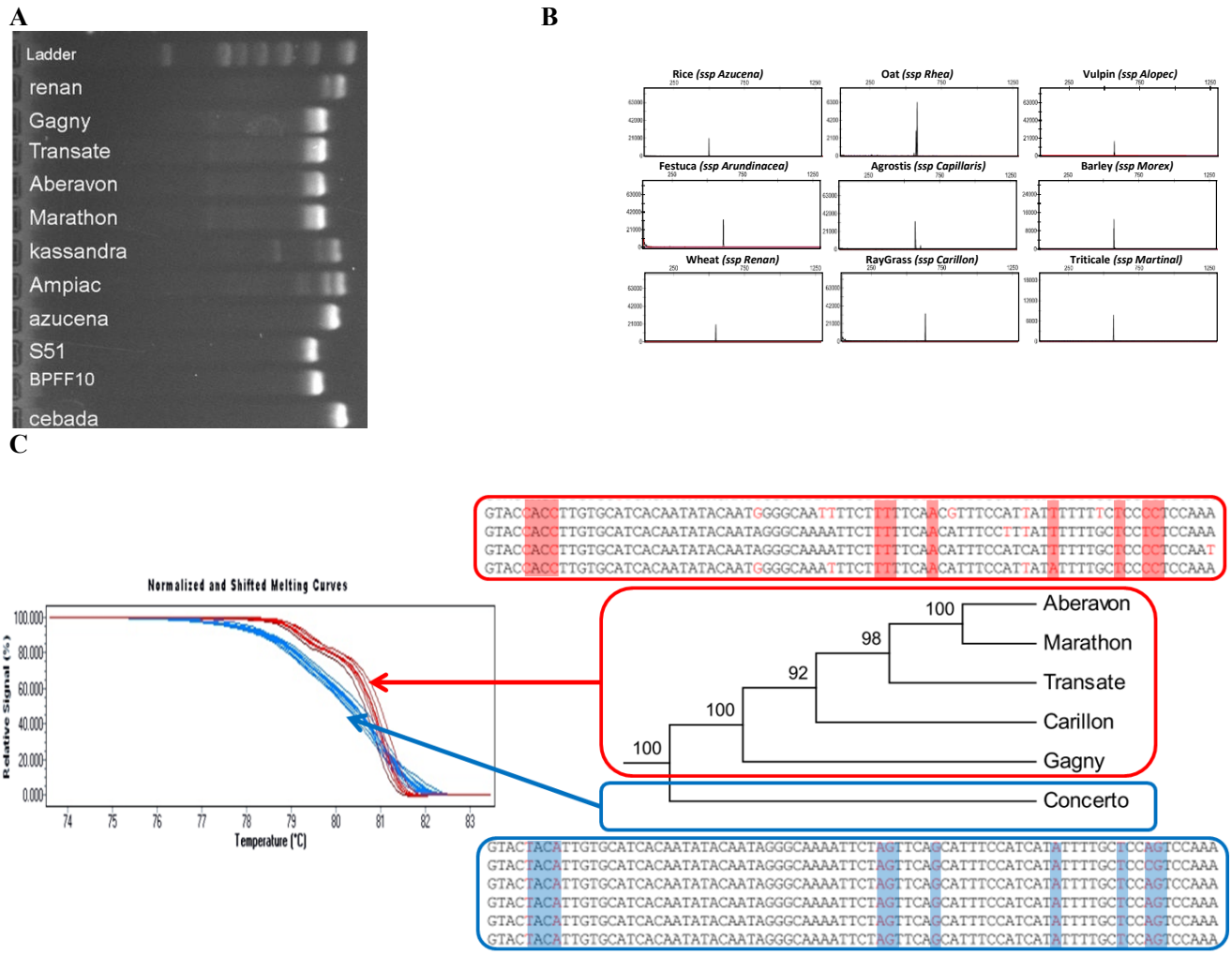
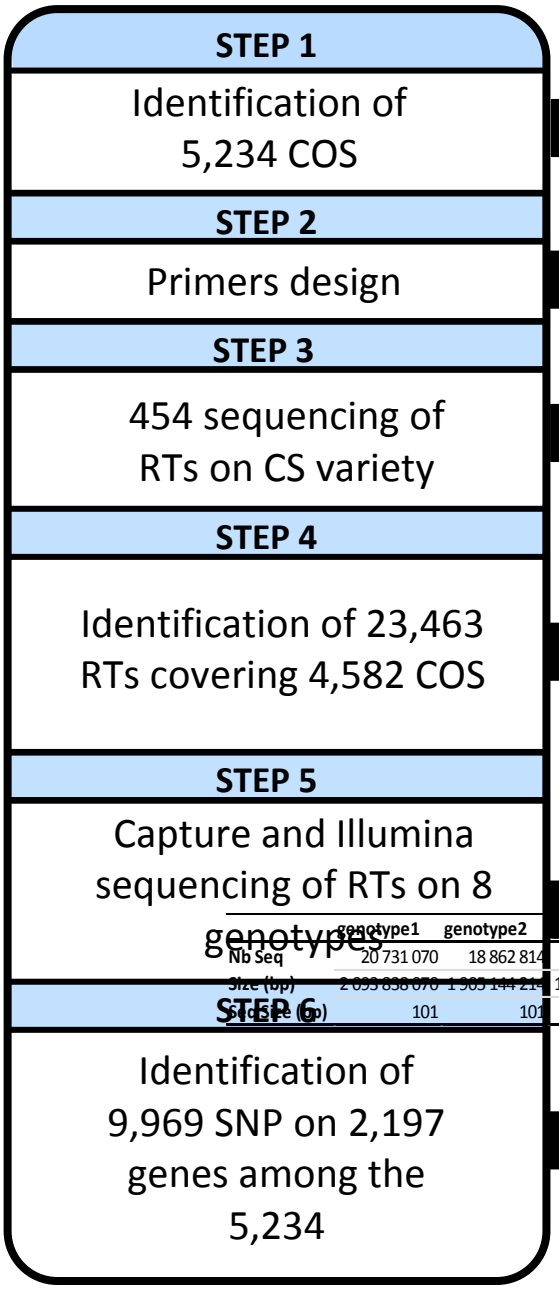


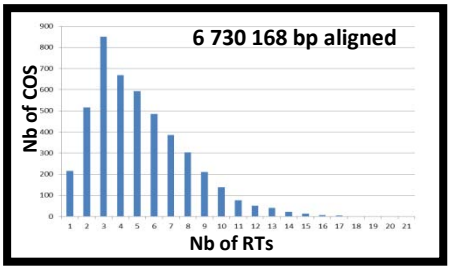
Figure S6: ***De novo COS sequencing in wheat***. The six steps followed for COS sequencing and SNP detection are illustrated at the left as follows: **step 1** for COS identification, **step 2** for primer design, **step 3** for reference transcripts (RTs) sequencing, **step 4** for homoeologous RTs identification, **step 5** for RTs capture and sequencing in 8 genotypes, **step 6** for SNP calling. Number of reads, sequence coverage, sequence distribution, number of contigs, number of SNPs are mentioned are the right for each step.



Wheat syntenome

COS finder

	Reads	Contigs
Nb Seq	1 271 355	814 914
Size (bp)	402 877 250	278 592 453
Seq Size (bp)	317	342



	genotype1	genotype2	genotype3	genotype4	genotype5	genotype6
Nb Seq	20 731 070	18 862 814	19 410 677	17 617 897	17 399 574	25 366 139
Size (bp)	2 093 838 070	1 905 144 214	1 960 478 377	1 779 407 597	1 757 356 974	2 561 980 039
Seq Size (bp)	101	101	101	101	101	101

SNP calling strategy (Figure S6)

Figure S7: ***HomoeoSNP vs SNP calling through sequence coverage criterion.*** **(Top) SNP identification pipeline.** Workflow illustration of the SNP identification pipeline consisting in reads alignment of the COS reference transcripts (RTs) for each genotype (Genotype 1 to 8) investigated; and the identification for each covered nucleotide of variants (red and green vertical bars) associated with a position and sequence coverage (for Forward, Reverse or Assembled F and R sequences). **(Bottom) SNP calling criteria.** Illustration at the bottom left of a homoeoSNP (T/A considered as 'low quality') detection where the sequence coverage is <0.2 and of an SNP (T/A) detection where the sequence coverage is >0.25. Based on the sequence coverage associated for each nucleotide (bottom right), six SNP calling classes are defined as illustrated from class A to class F (bottom middle).

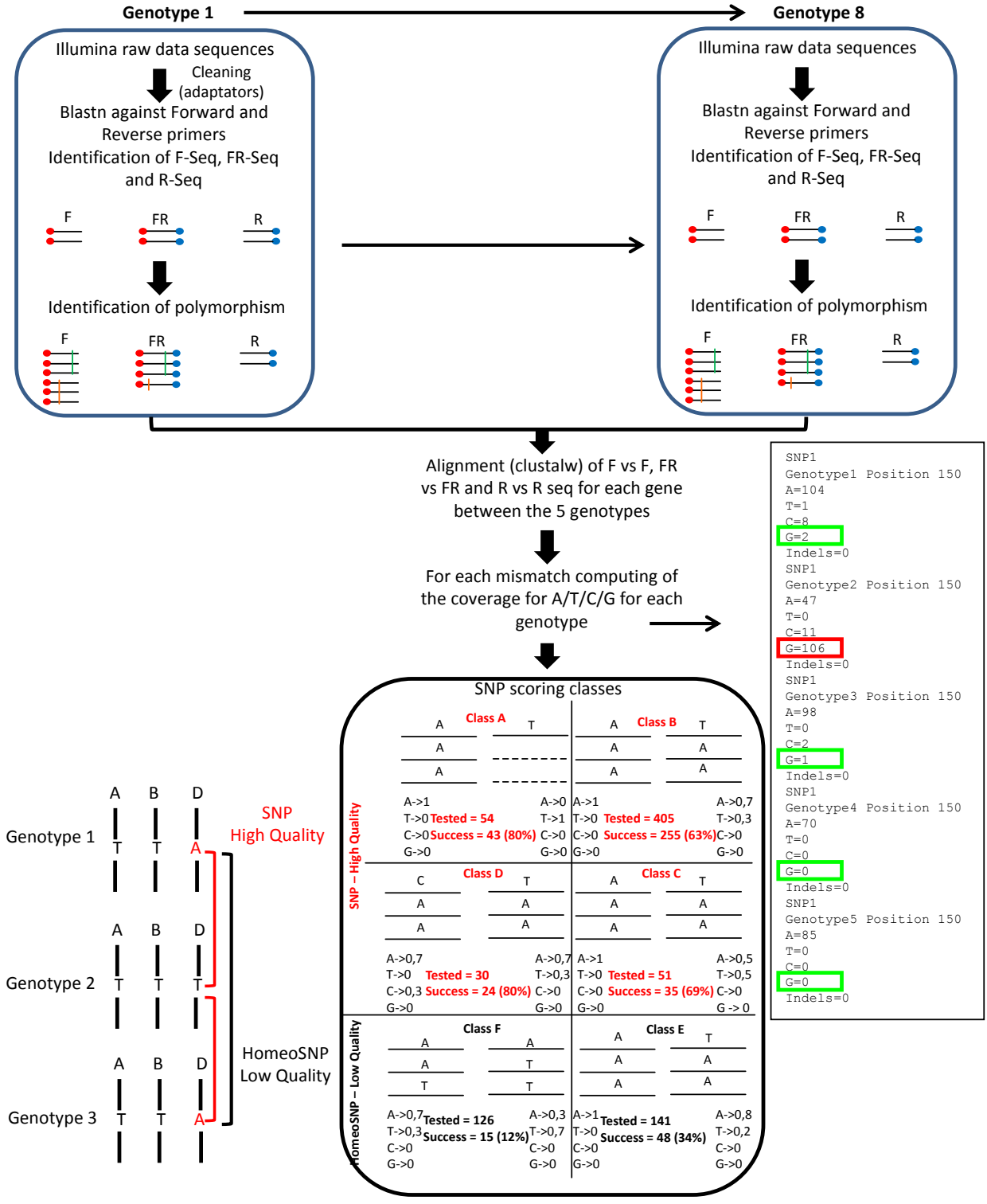


Figure S8: ***Characterization of wheat homoeologs (A, B and D hamoeoalleles).*** Sequence similarity curves obtained in aligning the wheat homoeologs defined from Brenchley *et al.* 2012, top distribution), and detailed in Figure 2B are illustrated as grey connecting line on the wheat circle (left). Alignments of A *vs*. B (blue curve), A *vs*. D (red curve), B *vs*. D (green curve) illustrate sequence similarity differences (highlighted with red vertical arrows) observed between homoeologs. In order to assign the 23,463 RT sequences on homoeologous groups (A, B and D) we used the previous sequence similarity curves established for known assigned COS sequences (middle distribution) as a control to extrapolate the number of A, B and D homoeologs present in the total set of 23,463 COS sequences (bottom distribution). A binomial test (=1/2) has been performed to test the observed retention of A, B and D homoeologs (blue bars) characterised within the total set of 23,463 RTs, using the previous sequence similarity curves.

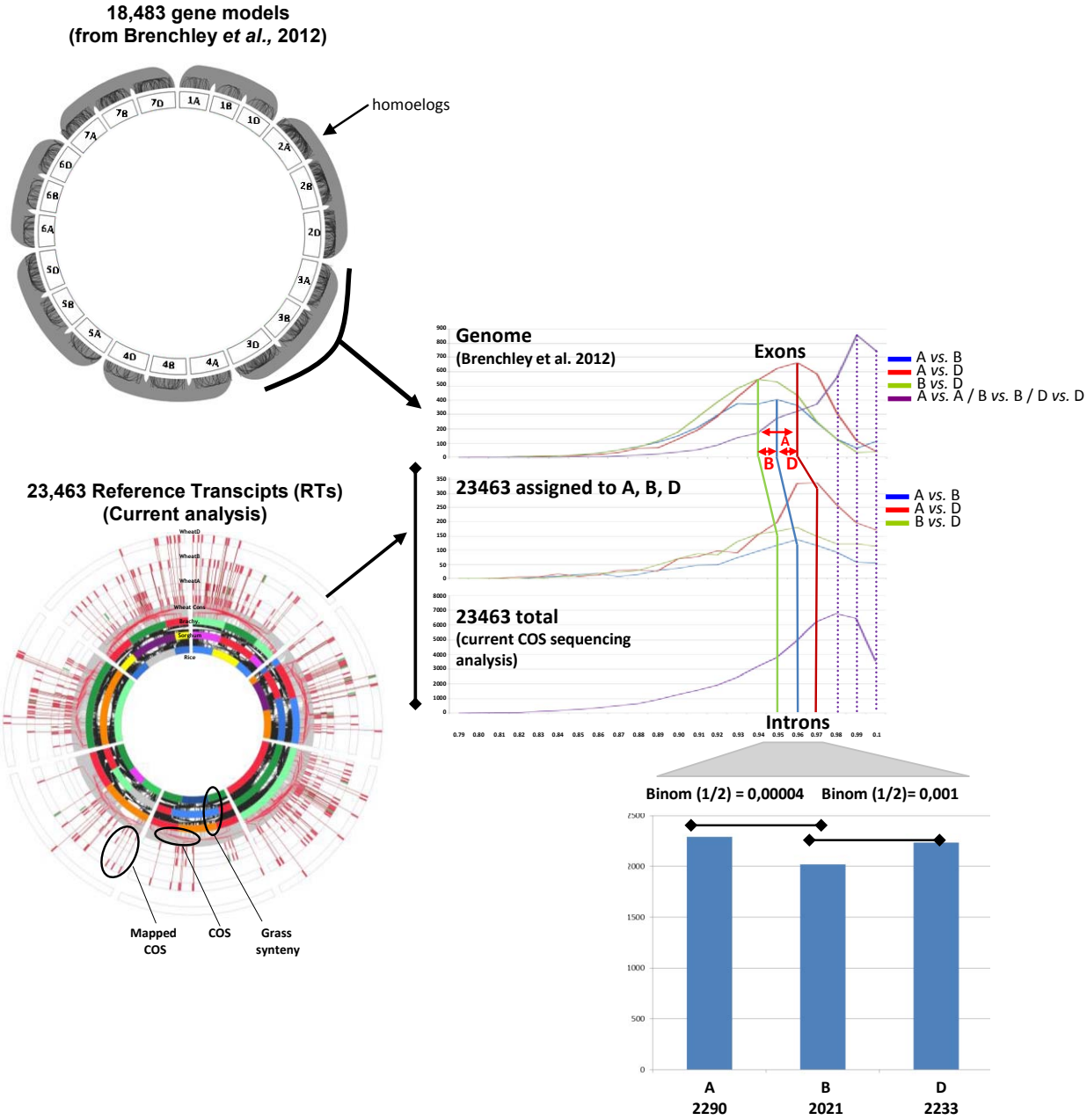
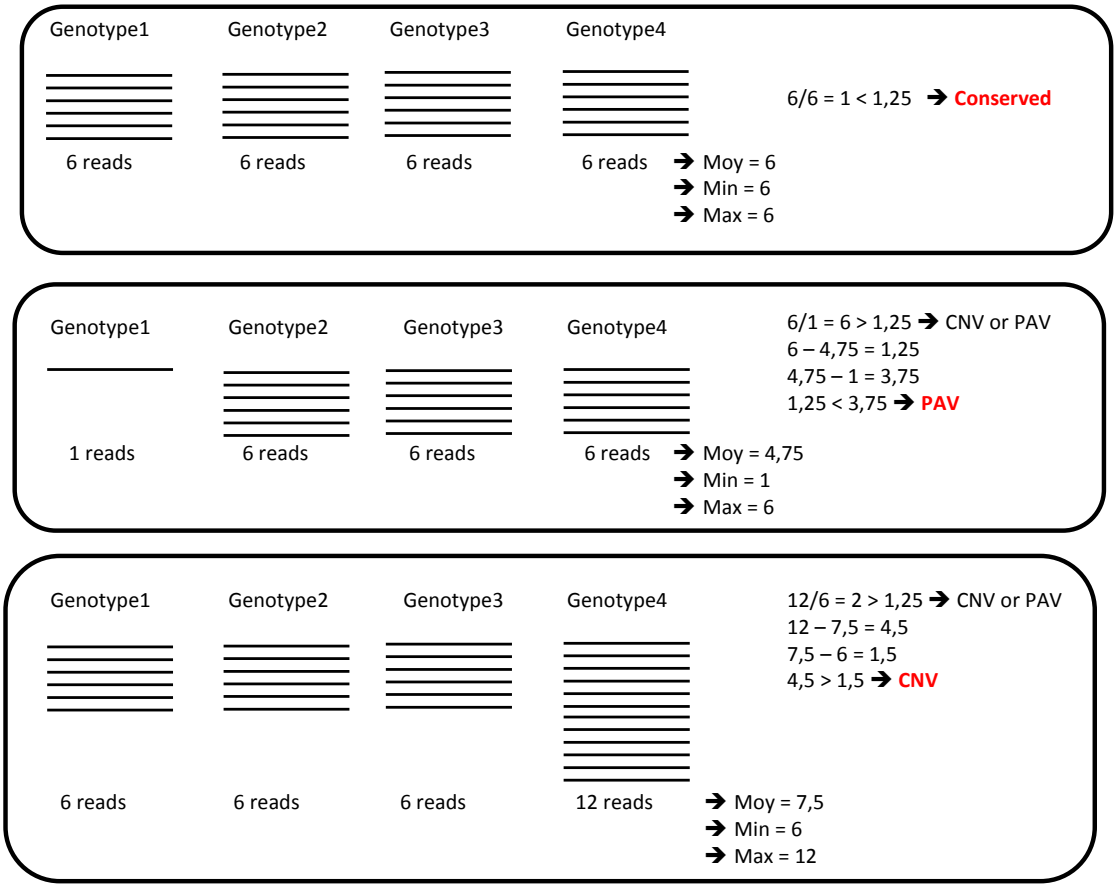


Figure S9: ***Identification of PAVs and CNVs***. **A-** illustration of the strategy used in defining NSVs, PAVs, CNVs based on sequence coverage criteria. **B-** Three distinct COS marker are used to illustrate (*i*) association with sequences obtained for all the eight (1 to 8) investigated (COS-5368, NSV observed in 80% of the cases), (*ii*) absence of sequences (COS-542, PAV observed in 6% of the cases), (*iii*) presence of extra sequence copies (COS-3070, CNV observed in 14% of the cases). The percentage of NSVs (SNPs or InDels), PAVs and CNVs observed for the 5 234 COS investigated are mentioned at the left end side.

A



B

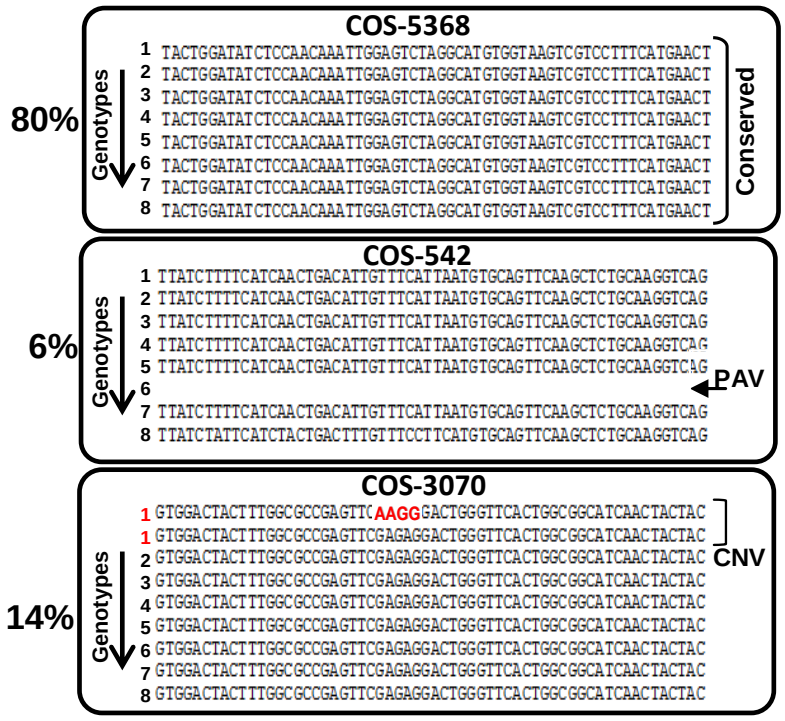
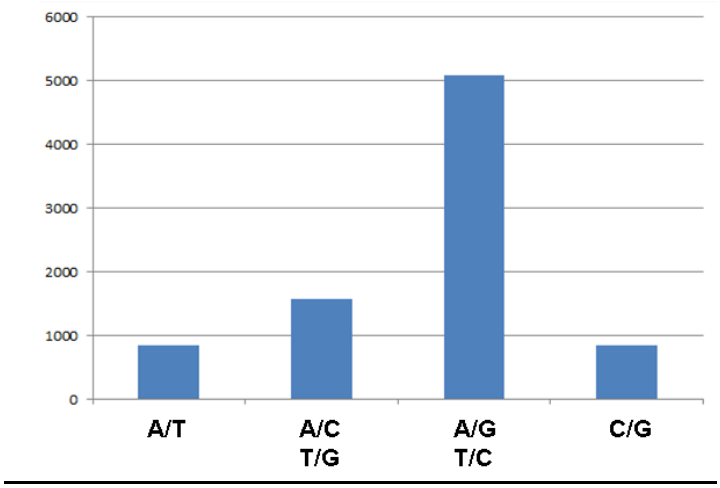


Figure S10: ***COS-SNP and InDels typology in wheat.*** **A-** Distribution of COS-SNPs type (ATGC) in wheat. The Y-axis represents the number of COS and the X-axis represents the different sequence substitutions. **B-** Distribution of COS-InDel sizes in wheat. The Y-axis represents the number of InDels and the X-axis represents the InDel sizes (from 1 to 10 base deletions).

A



B

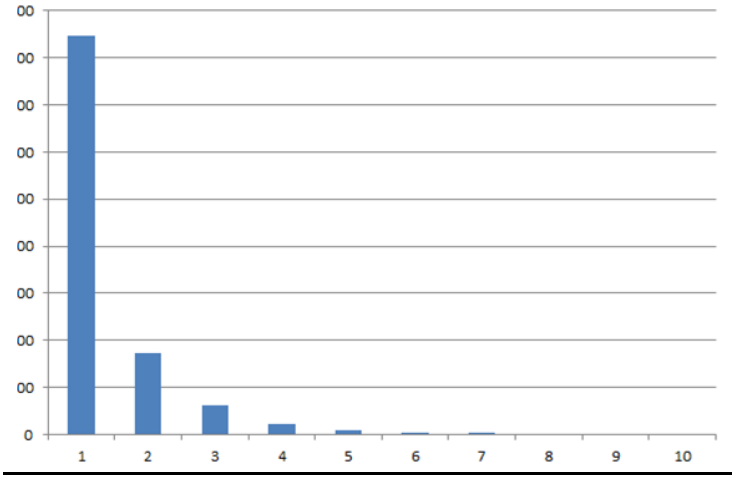


Figure S11: ***Genome-wide distribution of the COS-SNP diversity in wheat***. Heat-map of SNP diversity in wheat (chromosomes w1 to w7) where the number of characterized SNP per COS (Y-axis) is plotted along the chromosomes according to the virtual gene order positions in wheat. Centromeric ('cent.') regions (horizontal plain black lines) are shown as well as synteny breakpoints (vertical red dotted lines) defined using rice genomes as the closest representative of the grass ancestral genome.

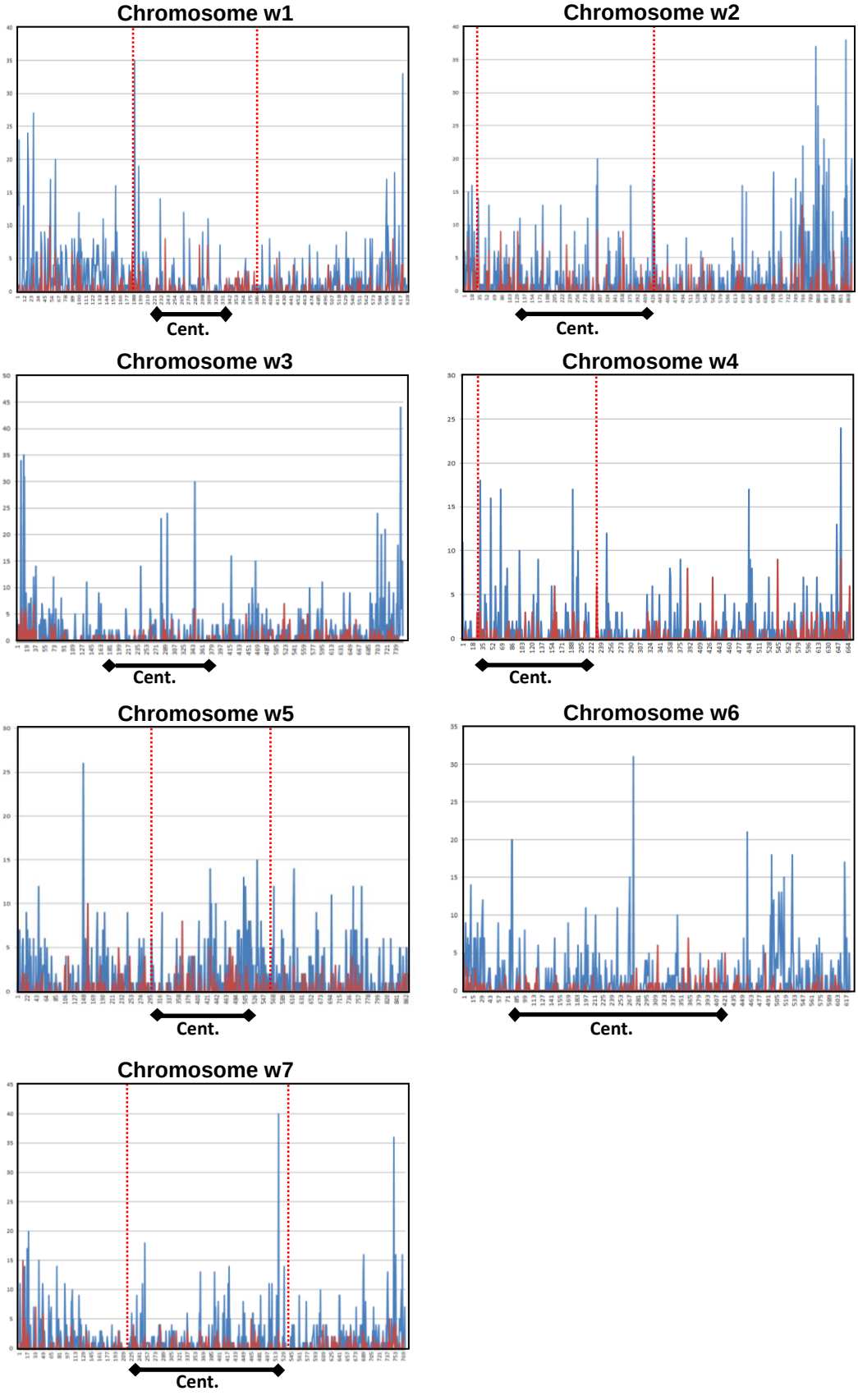
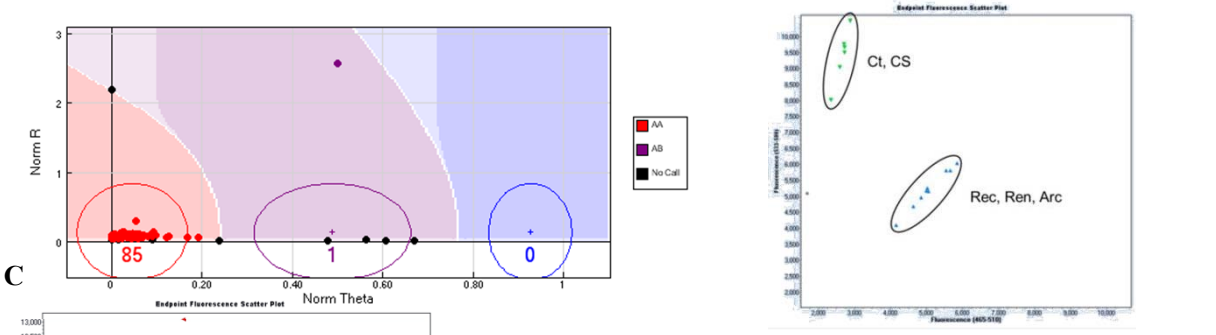


Figure S12: ***False COS-SNP origins.*** **A-** Four COS-SNP markers (5511-5598-5328-5820) are illustrated to unravel bias in methodological detection (Illumina vs. Kaspar) for SNP (ATCG associated with a sequence coverage score (highlighted in red). **B-** Methodology detection difference between Illumina (left) *vs* KASpar (right) is illustrated below with the COS-5328, a high quality (SNP scoring class A) SNP [T/C], showing no polymorphism with Illumina genotyping approach (B-left) and polymorphism between varieties Courtot/Chinese Spring and Renan/Recital/Arche using KASpar methodology (B-right). **C**- The second origin of false or non validated *in silico* SNPs is related to the presence of homoeoSNPs even with high quality criteria (SNP scoring class A) for the considered SNP variant. The COS-5820 sequence variant (C-right) did not show any polymorphism using the Illumina approach on the tested genotypes but show polymorphism (C-left) on cytogenetic material, proof of the presence of an homoeoSNP. **D**- The third origin of false or non validated *in silico* SNPs consists in identical high quality (SNP scoring class A) SNPs that are either validated by any of the two genotyping methods (D-top) or systematically monomorphic, *i.e.* false SNP, even with an average of 50 reads coverage supporting the position and nature of the putative SNP.

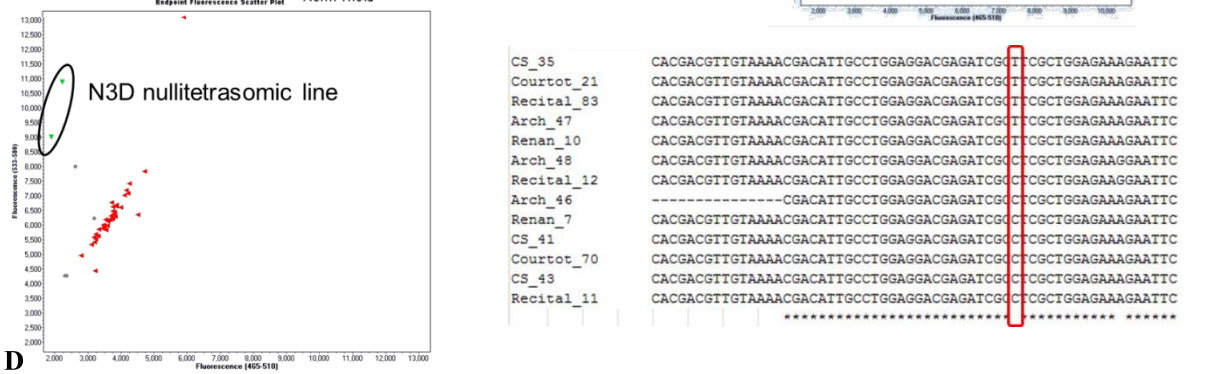
A

			Arch				Courtot				Recital				Renan				CS			
	illumina	kaspar	A	T	C	G	A	T	C	G	A	T	C	G	A	T	C	G	A	T	C	G
COS-5511	no	no	0,00	0,00	1,00	0,00	0,00	0,46	0,54	0,00	0,00	0,41	0,59	0,00	0,00	0,33	0,67	0,00	0,00	0,37	0,62	0,00
COS-5598	yes	yes	0,45	0,00	0,00	0,54	0,43	0,00	0,00	0,57	0,36	0,00	0,00	0,64	0,47	0,00	0,00	0,53	0,00	0,00	0,00	1,00
COS-5328	no	yes	na	na	na	na	na	na	na	na	0,00	1,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	1,00	0,00
COS-5820	no	no	0,85	0,00	0,00	0,15	0,91	0,00	0,00	0,09	0,74	0,00	0,00	0,26	1,00	0,00	0,00	0,00	0,87	0,00	0,00	0,13

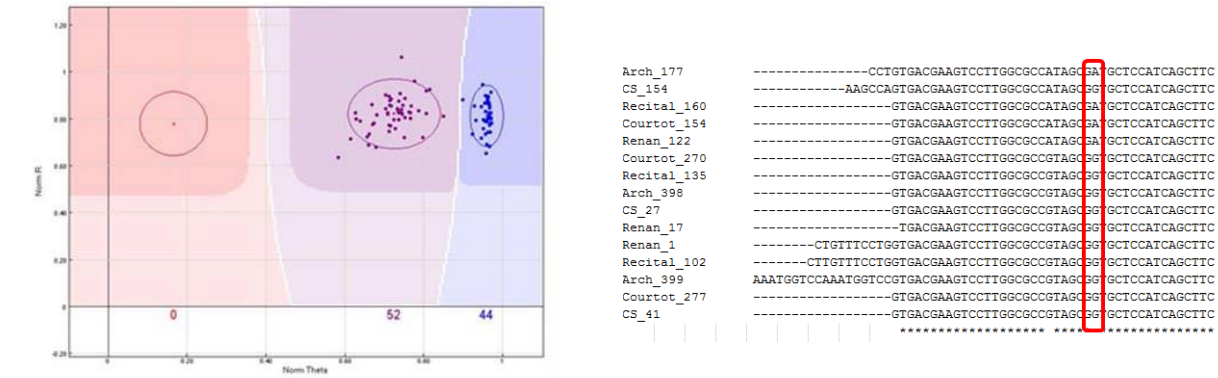
B



C



D



<u>Supplemental Figure 12</u> (end)

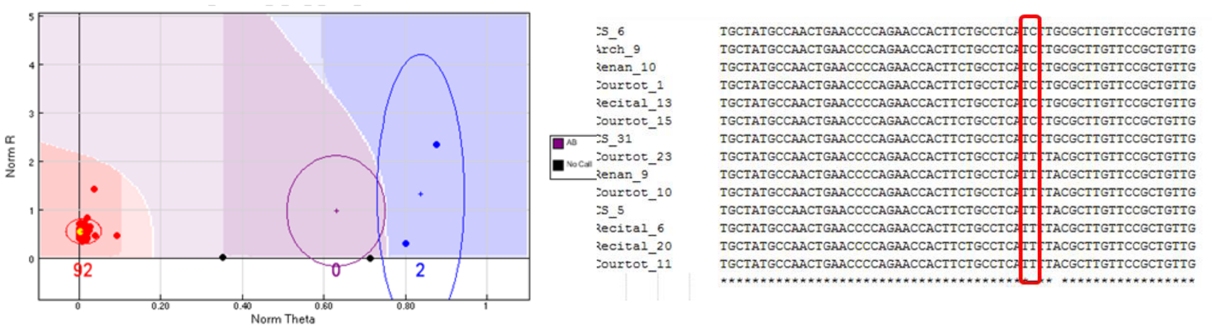


Figure S13: ***Validation of COS computed gene order***. The mapped COS-SNP markers on the wheat chromosome group 1 (W1) are compared to the location of the orthologous counterparts in rice (chromosome 5), *Brachypodium* (chromosome 2) and sorghum (chromosome 9). The percentage of conserved gene orders between wheat subgenomes 1A, 1B and 1D and the reference sequenced grass genomes are indicated at the bottom of the chromosomes.

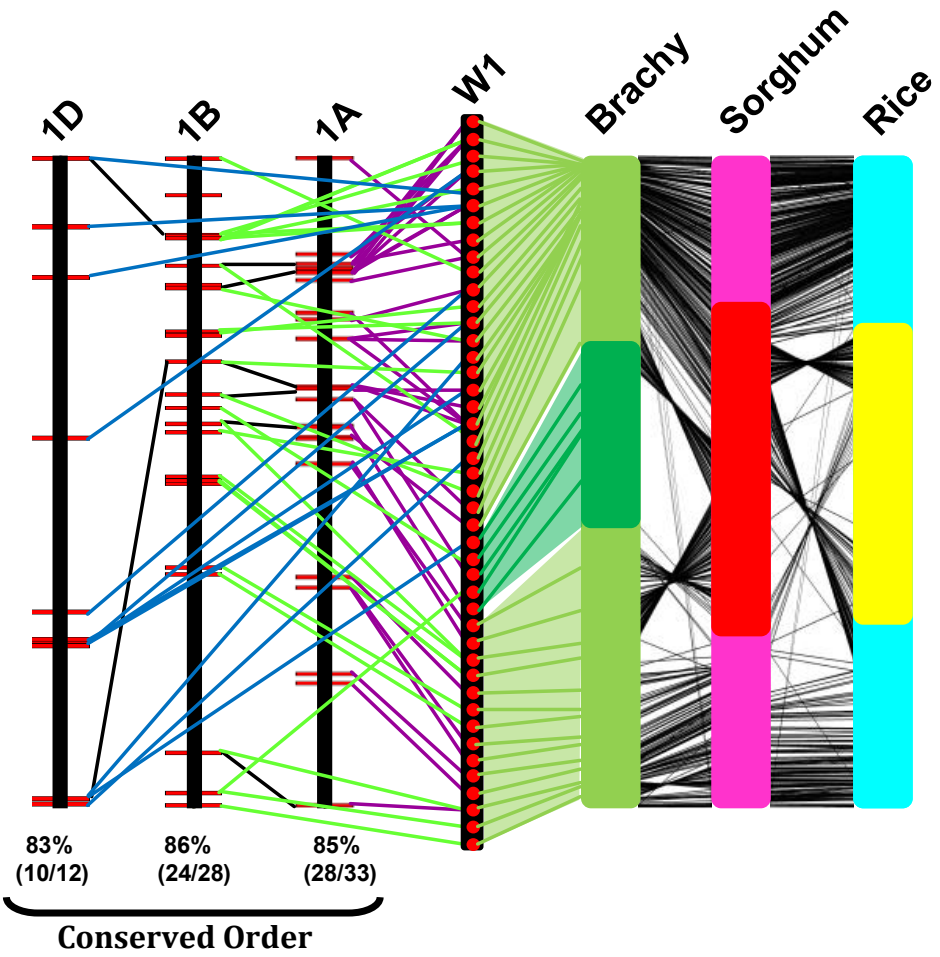


Figure S14: ***Detailed characteristics of the WCGM 2013.*** **A-** Number of markers (RFLP, AFLP, STS, SSR, DArTs and COS) per chromosome in bread wheat (table lines). **B-** Distribution of markers (RFLP, AFLP, STS, SSR, DArTs and COS as color code on the Y-axis) per chromosome in bread wheat (X-axis).

A

MARKERS

	other (ge	RFLP	AFLP	STS	SSR	DArT	COS	total
1A	48	99	37	18	128	71	38	439
1B	75	129	40	24	143	123	31	565
1D	41	81	41	7	98	39	13	320
2A	37	114	31	13	129	44	28	396
2B	37	98	18	19	113	114	37	436
2D	29	94	31	12	148	38	5	357
3A	40	86	34	25	100	63	27	375
3B	89	90	58	43	117	130	54	581
3D	22	67	10	2	119	7	10	237
4A	40	80	79	9	118	63	8	397
4B	28	56	20	1	81	30	11	227
4D	13	48	0	0	86	6	0	153
5A	32	81	40	4	108	21	7	293
5B	46	85	58	19	119	64	25	416
5D	25	46	12	0	113	4	1	201
6A	44	74	42	4	87	57	22	330
6B	68	80	44	18	89	95	23	417
6D	24	62	11	0	86	3	8	194
7A	66	89	52	17	118	135	29	506
7B	58	88	49	27	124	125	20	491
7D	28	40	5	0	91	14	11	189
	890	1687	712	262	2315	1246	408	7520

B

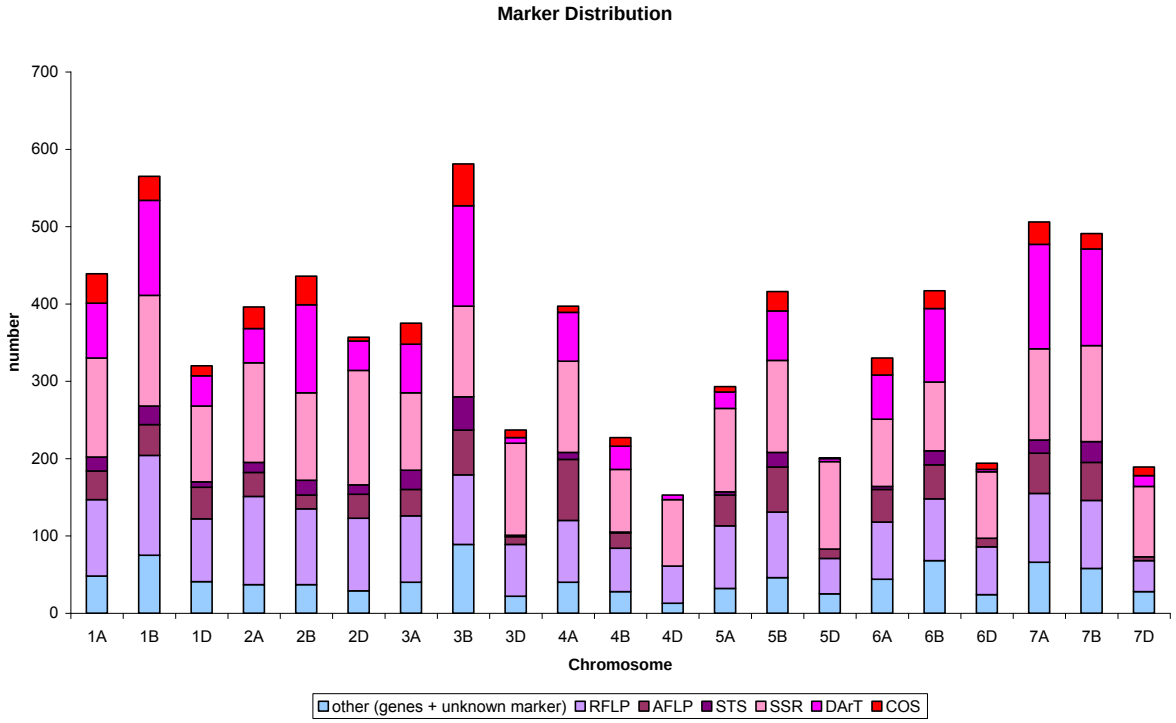
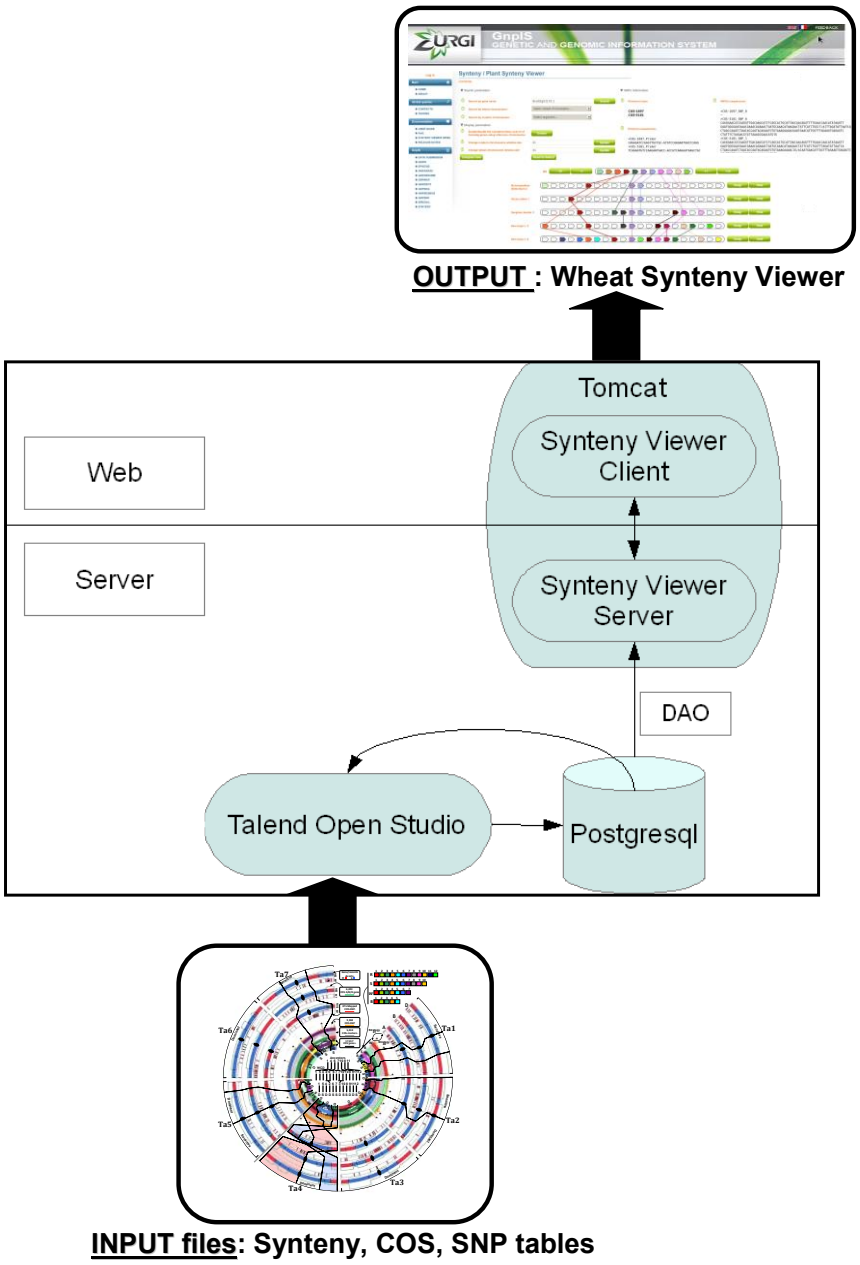


Figure S15: ***Wheat Synteny Viewer characterization flow chart***. Schematic representation of the algorithm used to develop the Wheat Synteny Viewer tool available at http://urgi.versailles.inra.fr/synteny-wheat.



SUPPORTING EXPERIMENTAL PROCEDURE

WHEAT SYNTENOME PROTOCOL - COS IDENTIFICATION- The presence in protein or nucleotide sequences of short, highly-conserved motifs can easily lead to artefactual attribution of orthologous or paralogous relationships between genes. To increase the significance of inter-specific coding sequence (CDS) alignments for inferring evolutionary relationships between genomes, we used two sequence alignment parameters for BLAST analyses (either nucleic or protein-based), which take into account not only similarity but also the relative lengths of the sequences (Salse *et al.,* 2009): CIP for Cumulative Identity Percentage and CALP for Cumulative Alignment Length Percentage. The Cumulative Identity Percentage, CIP [(∑ *nb* ID by (HSP/AL) x 100], corresponds to the cumulative percent of sequence identity observed for all the HSPs divided by the cumulative Aligned Length (AL) which corresponds to the sum of all HSP lengths. The Cumulative Alignment Length Percentage, CALP [AL / Query length], is the sum of the HSP lengths (AL) for all HSPs divided by the length of the query sequence. With these parameters, BLAST produces the highest cumulative percentage identity over the longest cumulative length thereby increasing stringency in defining conservation between two compared genome sequences (Salse *et al.,* 2009). **CAR IDENTIFICATION-** The reconstruction of ancestral karyotypes (*i.e.* protochromosomes) was obtained by computing common intervals of conserved blocks between two genomes (*i.e.* derived from the validated orthologous genes/blocks) and within a single genome (*i.e.* derived from the validated paralogous genes/blocks) into Contiguous Ancestral Regions (CAR) (Salse 2012). Briefly, chromosomal blocks that are duplicated in two different genomes but located at orthologous positions when comparing the two genomes are considered as (1) unique in the ancestor (*i.e.* CAR) and (2) deriving from a shared pre-speciation duplication event. In contrast, a chromosomal block that is duplicated in one genome but not identified as duplicated at orthologous positions when comparing two genomes is considered as (1) a species-specific duplication and (2) deriving from a post-speciation duplication event. The same approach is applied for any type of rearrangements including inversions and translocations. From the identified CARs, the most likely evolutionary scenario is proposed on the following assumptions: (1) ancestor modelling is based on duplications (or any shuffling events) found at orthologous positions between modern species, and thus considered as ancestral, (2) evolutionary history is based on the smallest number of shuffling operations (including inversions, deletions, fusions, fissions, translocations) that explain evolution from the ancestral genome to modern karyotypes.- **COMPUTATIONAL GENE ORDER IN WHEAT** - Several methods, such as InferCARs (Ma *et al.,* 2006), MGRA (Alekseyev *et al.,* 2009), ANGES (Jones *et al.,* 2012), are available to order, within previously described ancestral protochromosomes, genes conserved at orthologous positions between modern genomes. Among them only ANGES (Jones *et al.,* 2012) allows non universal genes and losses so it is the only way to infer gene orders for the AGK. For AGK intermediate ancestors, MGRA uses a model with a limited set of possible events while plant evolution seems more complex. ANGES, similar to InferCARs in principle, is more general as it computes ancestral adjacencies and intervals (only adjacencies for InferCARs), and has been tested on a wide range of kingdoms: plants, animals, bacteria, fungi (in contrast to InferCARs only tested on mammals). ANGES was used in the current analysis to produce ancestral gene order for AGK (n=7) and the intermediate ancestor (n=12). **D-S BLOCKS IDENTIFICATION** - Ancestral duplicated regions (paleo-S and paleo-D) conserved in grasses (rice, maize, sorghum, *Brachypodium*) were transferred on wheat chromosomes based on the synteny relationships. Paleo-S/paleo-D (taking into account all the seven duplicated blocks as well as excluding A11-A12 ancestral duplication associated with large gene conversion events, Jacquemin *et al.,* 2009) and A, B and D subgenomes were then compared for their retention of ancestral genes (number of retained genes consistent with the number or orthologous relationships also referenced as number of COS). For each pair of ancestral/recent duplicated chromosomes we characterized the number of retained ancestral genes (*i.e.* genes that are conserved between the investigated grass species) and defined dominant (highest number of retained genes) and sensitive (lowest number of retained genes) chromosomal blocks. In order to validate the observed partitioning and the variance of gene retention/deletion between ancient and recent D vs. S blocks, we performed paired t-test and binomial test (p=1/2) between each duplicated blocks. If the obtained p-value is lower than 0.05, we considered significant differences (in number of orthologs, number of COS, SNP/COS ration, cM/marker ratio, Number of C-band) between D/S or A/B or B/D genomic blocks.

WHEAT COS SEQUENCING PROTOCOL - COS-PRIMER DESIGN - Wheat EST-contigs exon structures were identified through rice-sorghum-maize-*Brachypodium*/wheat sequence alignments, as conserved HSPs correspond to exons. Precise exon/intron boundaries identified (*i.e.* HSP boundaries) for any considered wheat mapped ESTs associated with a rice-sorghum-maize-*Brachypodium* ortholog was

considered in order to define two values, *i.e.* Ir and Er. Ir (for Included region) and Er (for Excluded region) are associated with any Intron position (Ii) within Wheat EST-contigs aligned with a rice sequence: Er = (I_i-25) to (I_i+25). This region corresponds to 50 nucleotides centred on the predicted intron position within the wheat EST sequence. Ir= (I_{i-1}+10) to (I_{i+1}-10). This region corresponds to the 2 exons spanning the predicted intron position within the wheat EST sequence. The precise sequence region corresponding to Ir-Er is provided to Primer 3 package to select primer pairs on exons for intron amplification with the following parameters suitable for detection on Applied Biosystems (ABI) capillary sequencer: (i) Primer size (20 to 25 mer as default parameters), (ii) Amplicon size (between 250-800 bp as default parameters), (iii) Tm (between 57-63 as default parameters), (iv) GC clamp (equal to 2, *ie* a G or C at the 5' extremity as default parameters), (v) GC percentage (50% as default parameters), Quraishi *et al.,* (2009). The previous COS primer design strategy is automatically accessible through the COS-*finder* software available at http://urgi.versailles.inra.fr/synteny-wheat. **COS-SNP SEQUENCING AND CLUSTERING - DNA extraction** - Plants were grown in a greenhouse and DNA was extracted from leaves of 3 week-old seedlings according to a described CTAB protocol (Dobrovolskaya *et al.,* 2011). DNA were adjusted at 50 ng/µl after quantify by infinite® 200 plate reader (Tecan) using Quant-iT™ Picogreen® dsDNA Assay Kit according to the Molecular Probes (Invitrogen) procedure. **COS amplification and sequencing (454-Roche) -** In order to obtain sequence with high sequence coverage of each COS, different procedures were tested in order to normalize the PCR product quantities with beads purifications, dilution, picogreen quantification, and concentration adjustments. Homogeneous PCR fragments were produced in a total volume of 10 µl from 20 ng of genomic DNA (30 ng) with the GoTaq® Colorless Master Mix (Promega). Each step (amplification, pooling) was done in the same way by robotic pipetting to obtain less than 10% of non amplification. The average sizing of the COS is 410 pb with around 30% of variations. The final PCR products were quantify with fluorescent Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen), pooled together and purified by QIAquick PCR Purification Kit (Qiagen) according to the manufactory protocol. Library generation for the 454 FLX sequencing was carried out according to the manufacturer's standard protocols (Roche/454 life sciences, Branford, CT 06405, USA). Briefly, the bulk of PCR products were end-polished and the 454 A and B adaptors that are required for the emulsion PCR and sequencing were added to the ends of the fragments by ligation. The resulting fragment library was sequenced as shotgun protocol on the picotiterplate (PTP) on the GS FLX using the Roche/454 Titanium chemistry and protocols. **Sequence assembly and SNP/InDel/PAV/CNV scoring -** All reads of each genotype were aligned against the 23,463 reference sequences with BLAST. Coverage and proportion for all nucleotides (A, T, C, G, InDels) for each genotype at each position on reference sequences was calculated. In order to identify SNP, the proportion of each base was compared between each genotypes. When the total coverage at a position was smaller than 15, the considered genotype was excluded to the identification SNP scoring procedure. Each variation identified was classified with a quality score to separate *in silico* SNP (HighQuality) and homeo-SNP (LowQuality). SNP (high quality) are identified according to two types, on the one hand presence/absence polymorphism which corresponds to the absence of one base in a genotype (proportion ~ 0%) and the presence of this base (proportion > 30%) in another genotype. On the other hand, SNP with significant divergence of nucleotide proportion have been identified, these SNP correspond to a proportion of a nucleotide close to 100% in a genotype and a proportion included between 25% and 75% in another genotype. All variations that not correspond to these two cases are considered as homoeoSNP. To identify CNV and PAV between genotypes, the maximal number of reads aligned along a region for each gene was calculated. If the maximal number of reads among the 8 genotypes for a gene is 2.25x the minimal number of reads, a CNV or a PAV was considered. Then if average of the number of reads among the 8 genotypes for a gene is closer than the maximal number of reads the PAV was considered otherwise CNV was considered. SNP calling method consists in scoring the proportion of A, T, C, G and as well as deletion from the aligned reads (Solexa reads from eight genotypes) for each position of the reference sequence (454 sequence assembly from bread wheat cv. Chinese Spring). If a variation between two genotypes is detected, differential nucleotide coverages are compared between genotypes to identify if this variation is scored as homoeoSNP or SNP (Figure S2). If the absence of a given nucleotide is observed in a considered genotype while its sequence coverage is observed to be >25% in another, a SNP is scored as 'high quality' (SNP scoring classes A to D), otherwise a homoeoSNP (SNP scoring classes E to F) is considered as illustrated in Figure S5. The SNP/COS ratio then used to compare ancestral and recent D *vs.* S blocks.

SNP GENOTYPING PROTOCOL - *SSCP* - COS primers were synthesized with 5' extensions in order to facilitate the labelling procedure at low cost: forward primer with the CACGACGTTGTAAAACGAC

sequence extension and reverse primer with the CAGGAAACAGCTATGACC sequence extension. PCR fragments were produced in two steps. In a total volume of 15 µl, genomic DNA (30 ng) was first amplified with the following PCR mix: 10 mM Tris-HCL, 3,1 mM MgCl2, 50 mM KCl, 0.001% gelatine pH 8.3, 5% glycerol, 400 µM dNTP, 0.4 µM forward and reverse primers, 0.2U Taq polymerase (Qiagen). This PCR product was diluted (1/10) and re-amplified with the same PCR mix including 0.2 µM of each labeled primers (6-FAM and NED, Applied Biosystems) in a final volume of 15 µl. 2 µl of the PCR product was then diluted (1/10) and pooled with 0.2 µl of 900 bp MegaBace ET900-R Size Standard (GE Healthcare), 0.2 µl of 0.3 N NaOH and 9 µl HI-Di Formamide (Applied Biosystems). Fragments were separated by capillary electrophoresis on ABI3100 (Applied Biosystems) in 50 min with a 36 cm capillary. The running polymer consists in 1x of running buffer, 5% Genscan Polymer (Applied Biosystems), 10% glycerol. Samples were denatured during 2 minutes at 95° C and 10 min in ice. The sample buffer consists in 1x of running buffer and 10% glycerol. After denaturing, the samples were injected at 2.5 kV during 50 seconds and separated at 18, 25, 35° C and 15 kV. Data were analysed using GeneMapper 3.7 software. **ILLUMINA** – The SNP-harboring sequences were then submitted to Illumina for processing by Illumina® Assay Design Tool (ADT) and they were genotyped using the Illumina BeadArray platform and GoldenGate Assay following the manufacturer's protocol on 150 ng of genomic DNA. Analysis was performed using the genotyping module in the BeadStudio package (Illumina, San Diego, CA, USA) and clusters generated for each SNP locus by GenCall. **HRM-** Polymerase chain reactions (PCR) were done according to Roche protocol LightCycler® 480 Hight Resolution Melting Master. In a total volume of 15 µl, genomic DNA (10 ng) was amplified with the master mix, 3 mM $MgCl_2$ and 0.4 µM forward and reverse primers. In order to obtain homogeneous and comparable results the DNA was quantify with fluorescent Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen). Polymorphism analysis was performed in duplicates to eliminate false positives. We used the recommended PCR parameters on the LC480 system to establish a gene scanning assay with the melting master. Analyses were performed using LightCycler 480 software with gene scanning module. **SIZE POLYMORPHISM -** The DNA amplicons were analyzed in a protocol using Applied Biosystems® FAM Fluorochrome using 384-well plates. PCR (15 µl per well) with 30 ng of DNA were performed with AmpliTaq Gold® PCR Master Mix (Applied Biosystems) according to the protocol. The standard PCR program for amplifying DNA was a denaturation step at 95°C for 5 min, 35 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s; and a final extension at 72°C for 5 min. After electrophoresis on an ABIPRISM3100 sequencer, the data were analyzed using GeneMapper v4 software to obtein the final profil results. **SEQUENCING PROTOCOL** - Amplified PCR products were sequenced according to the manufacturer's instructions and sequenced in both strands with the forward and reverse COS primers. Sequencing reactions were performed using BigDye® Terminator v3.3 Cycle Sequencing Kit (Applied Biosystems) with 10 ng of DNA and 0.25 µl of Big Dye. Fragments migrations were performed on 3130XL Genetic Analyser (Applied Biosytems) with polymer POP7 and UltraSeq36_POP7_1 default run module settings. The PCR products were sequenced using 454 Life Sciences™ technology as described by Margulies *et al.* (2005) and the Genome Sequencer FLX systems methods manual with a shotgun DNA library. Each sample corresponding to one genotype provided a single library to generate the emPCR and more than 900 000 DNA beads that they were loaded in a large region size. **KASPAR** - The KASPar (KBioscience Ltd., Hoddesdon, UK) assay was used to stufy SNPs and homeoSNPs. All assay primer sets were designed by KBioscience (KASpar-By-Design) and assay screening and genotyping were performed on the LightCycler® 480 Real-Time PCR System (Roche Applied Science) with KASPar SNP reagent Mix and 2.5 ng of genomic DNA. Details of the method used can be found at http://www.kbioscience.co.uk/. Analysis was performed using LightCycler 480 software with end point genotyping module. **LNA® DUAL-LABELED FLUOROGENIC PROBES** - The ABI7900HT real time PCR machine (applied biosystems) has been used to perform single nucleotide polymorphism genotyping with the use of LNA probes obtained from Sigma–Aldrich and used as conventional TaqMan probes. Briefly, PCR (10 µl per well) with 30 ng of DNA were performed with JumpStart™ Taq ReadyMix (Sigma–Aldrich) according to the manufacturer protocol (0.4µM of probe and PCR program: 94°C for 2 min, then 40 cycles of 94°C for 15 s, 60°C for 2 min). Fluorescent data generated by cleavage of the dual-labeled probes were collected at the end of PCR and data analysis for allelic discrimination was conducted using the SDS Software 2.3 (applied biosystems). The polymorphic base was placed at a central location within the sequence of the oligonucleotide, and the Tm was about 60 °C. In the figure 2B the sequences of probes (COS-5368, see Figure S6) are atggagcATCgctAtgg and atggagcACCgctatgg (LNA nucleotides are marked in upper case).

WHEAT COMPREHENSIVE CONSENSUS GENETIC MAP PROTOCOL - SELECTION OF PUBLIC REFERENCE GENETIC MAPS - Three genetic maps are considered in the literature as reference genetic map in wheat. The first is the ITMI map (International Triticeae Mapping Initiatives, (Nelson *et al.,* 1995abc; Röder *et al.,* 1998ab). There are in total 2293 different molecular markers covering 3980.4 cM, resulting in a marker density of one every 1.735 cM. There is also an additional advantage in using this map as the deletion bins of Chinese spring (Qi *et al.,* 2004) are assigned (Sourdille *et al.,* 2004), http://wheat.pw.usda.gov/ggpages/SSRclub/GeneticPhysical/). The second possible choice is the genetic map constructed by R. Appels (Wheat Composite 2004 map) which is available at http**:**//wheat.pw.usda.gov/. This genetic map is a consensus genetic map of different mapping populations including Synthetic-W7984 x Opata85 (four studies), Arina x Forno, CD87 x Katepwa, CS x DH, Cranbrook x Halberd, Egret x Sunstar, and Sunco x Tasman. In total there are 3660 different markers on the genetic map covering 3121 cM, with an average distance of 0.85 cM between two markers. The third possible genetic map has been produced by D. Somers's group (Somers *et al.,* 2004). Four mapping populations, *i.e.* Synthetic-W7984 x Opata85 (68 RILs), RL4452 x AC Domain (91 DH lines), Wuhan x Maringa (93 DH lines), and Superb x BW278 (186 DH lines) have been integrated into a single consensus map, WCGM2013. **CONSTRUCTION OF THE WHEAT COMPOSITE GENETIC MAP (WCGM2013) -** Biomercator version 2.0 (Arcade *et al.,* 2004) has a graphical interface that allows the projection of different genetics maps with QTL obtained in independent genetic backgrounds, different traits, and different locations into a single genetic consensus map. A text file is necessary to describe all the genetic maps (marker name, position) and their associated QTL statistics (LOD score, R^2 percentage of phenotypic variation explained, Confidence Interval). Biomercator first integrates the independent genetic maps into a comprehensive map (with a specific map projection algorithm). As a consequence, we used the first function of Biomercator to compile two large genetic maps to create a dense Wheat Composite Genetic Map (WCGM2013) with all the markers available from the ITMI, Wheat Composite 2004 genetic maps, as well as the genetics maps obtained from the current COS-SNP mapping data [*Arche* x *Récita*l (A x Re), Laperche *et al.,* 2007; *Renan* x *Récital* (R x Re), Groos *et al.,* 2003, 2007, Quraishi *et al.,* 2011ab; *Courtot* x *Chinese Spring* (Ct x Cs), Perretant *et al.,* 2000, Charmet *et al.,* 2001]. The third available reference genetic map from Somers *et al.* (2004) has not been included in the construction of the WCGM2013 map as most of the markers available were present either in the ITMI or Wheat Composite 2004 genetic maps. A prerequisite for producing a comprehensive composite genetic map based the 5 previous specific maps was to eliminate inconsistent markers, *i.e.* markers located on non-identical positions between the five maps, so that they could not create discrepancies in the final comprehensive map. As a consequence, we used the MapInspect software (http://www.dpw.wau.nl/pv/pub/), to verify chromosome by chromosome the marker order between the two considered maps. All the inconsistent loci (non-collinear markers corresponding to large inversions) were thus discarded. The cM/ marker ratio then used to compare ancestral and recent D *vs.* S blocks.

REFERENCES CITED:

Alekseyev, M.A. and Pevzner, P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* **19**(5), 943-57.

Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A. and Joets, J. (2004) BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* **20**, 2324-2326.

Charmet, G., Robert, N., Perretant, M.R., Gay, G., Sourdille, P., Groos, C., Bernard, S. and Bernard, M. (2001) Marker assisted recurrent selection for cumulating QTLS for bread-making related traits. *Euphitica* **119**, 89-93.

Dobrovolskaya, O., Boeuf, C., Salse, J., Pont, C., Sourdille, P., Bernard, M. and Salina, E. (2011) Microsatellite mapping of Ae. speltoides and map-based comparative analysis of the S, G, and B genomes of Triticeae species. *Theor Appl Genet.* **123**(7), 1145-57.

Groos, C., Robert, N., Bervas, E. and Charmet, G. (2003) Genetic analysis of grain protein-content, grain yield and thousand-kernel weight in bread wheat. *Theor Appl Genet*. **106**, 1032-1040.

Groos, C., Bervas, E., Chanliaud, E. and Charmet, G. (2007) Genetic analysis of bread-making quality scores in bread wheat using a recombinant inbred line population. *Theor Appl Genet*. **115**, 313-323.

Jacquemin, J., Laudié, M. and Cooke, R. (2009) A recent duplication revisited: phylogenetic analysis reveals an ancestral duplication highly-conserved throughout the Oryza genus and beyond. *BMC Plant Biol*. **9**:146.

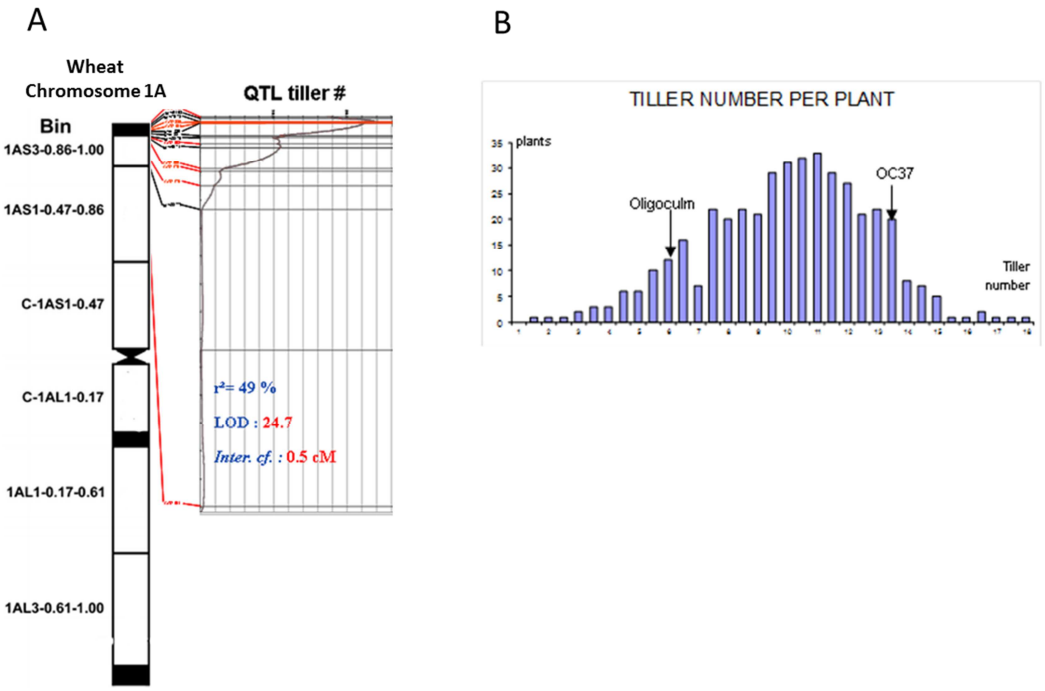
Jones, B.R., Rajaraman, A., Tannier, E. and Chauve, C. (2012) ANGES: Reconstructing ANcestral GEnomeS maps. *Bioinformatics* **28**(18), 2388-90.
Laperche, A., Brancourt-Hulmel, M., Heumez, E., Gardet, O., Hanocq, E., Devienne-Barret, F. and Le Gouis, J. (2007) Using genotype x nitrogen interaction variables to evaluate the QTL involved in wheat tolerance to nitrogen constraints. *Theor Appl Genet*. **115**, 399-415.
Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D. and Miller, W. (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res*. **16**(12):1557-65.
Nelson, J.C., Sorrells, M.E., Vandeynze, A.E., Lu, Y.H., Atkinson, M., Bernard, M., Leroy, P., Faris, J.D. and Anderson, J.A. (1995a) Molecular Mapping of Wheat: Major Genes and Rearrangements in Homoeologous Groups 4, 5, and 7. *Genetics* **141**, 721-731.
Nelson, J.C., Vandeynze, A.E., Autrique, E., Sorrells, M.E., Lu, Y.H., Merlino, M., Atkinson, M. and Leroy, P. (1995b) Molecular mapping of wheat. Homoeologous group 2. *Genome* **38**, 516-524.
Nelson, J.C., Vandeynze, A.E., Autrique, E., Sorrells, M.E., Lu, Y.H., Negre, S., Bernard, M. and Leroy, P. (1995c) Molecular mapping of wheat. Homoeologous group 3. *Genome*, **38**, 525-533.
Perretant, M.R., Cadalen, T., Charmet, G., Sourdille, P., Nicolas, P., Boeuf, C., Tixier, M.H., Branlard, G., Bernard, S. and Bernard, M. (2000) QTL analysis of bread-making quality in wheat using a doubled haploid population. *Theor Appl Genet*. **100**, 1167-1175.
Qi, L.L., Echalier, B., Chao, S. *et al.* (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics*, **168**, 701-712.
Quraishi Masood, M., Abrouk, M., Bolot, S. *et al.* (2009) Genomics in cereals: From genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. *Functional & Integrative Genomics* **9**(4), 473-84.
Quraishi, U.M., Murat, F., Abrouk, M. *et al.* (2011a) Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (*Triticum aestivum* L.). *Functional & Integrative Genomics* **11**(1), 71-83.
Quraishi, U.M., Abrouk, M., Murat, F. *et al.* (2011b) Cross-genome map based dissection of a nitrogen use efficiency ortho-metaQTL in bread wheat unravels concerted cereal genome evolution. *Plant J.* **65**(5), 745-56.
Roderc M.S., Korzunc V., Gillc B.S. and Ganalc M.W. (1998a) The physical mapping of microsatellite markers in wheat. *Genome* **41**, 278-283
Roder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. (1998b) A microsatellite map of wheat. *Genetics* **149**, 2007-2023.
Salse, J., Abrouk, M., Murat, F., Masood Quraishi, U. and Feuillet, C. (2009b) Improved standards and new comparative genomics tools provide new insights into grasses paleogenomics. *Briefings in Bioinf.* **10**(6), 619-30.
Salse, J. (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr Opin Plant Biol*. **15**(2), 122-30.
Somers, D.J., Isaac, P. and Edwards, K. (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum L.*). *Theor Appl Genet*. **109**, 1105-1114
Sourdille, P., Singh, S., Cadalen, T., Brown-Guedira, G.L., Gay, G., Qi, L., Gill, B.S., Dufour, P., Murigneux, A. and Bernard, M. (2004) Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). *Functional & Integrative Genomics* **4**, 12-25.

ANNEXE 3

Supplementary data de l'article Physical mapping of the Tiller INhibitor (TIN) gene in bread wheat (Chapitre 3)

Supplementary Figure 1: Tiller Inhibition Number QTL

(A) The Oligoculm x OC37 QTL (right, expressed in LOD score) localization on the chromosome 1AS3 (BIN 0.86-1.00) gnetic map (left). SSR markers are illuminated in black and COS markers in red. (B) Distribution of the tiller number (y-axis) in the Oligoculm x OC37population (x-axis with the parents Oligoculm and OC37 mentionned).



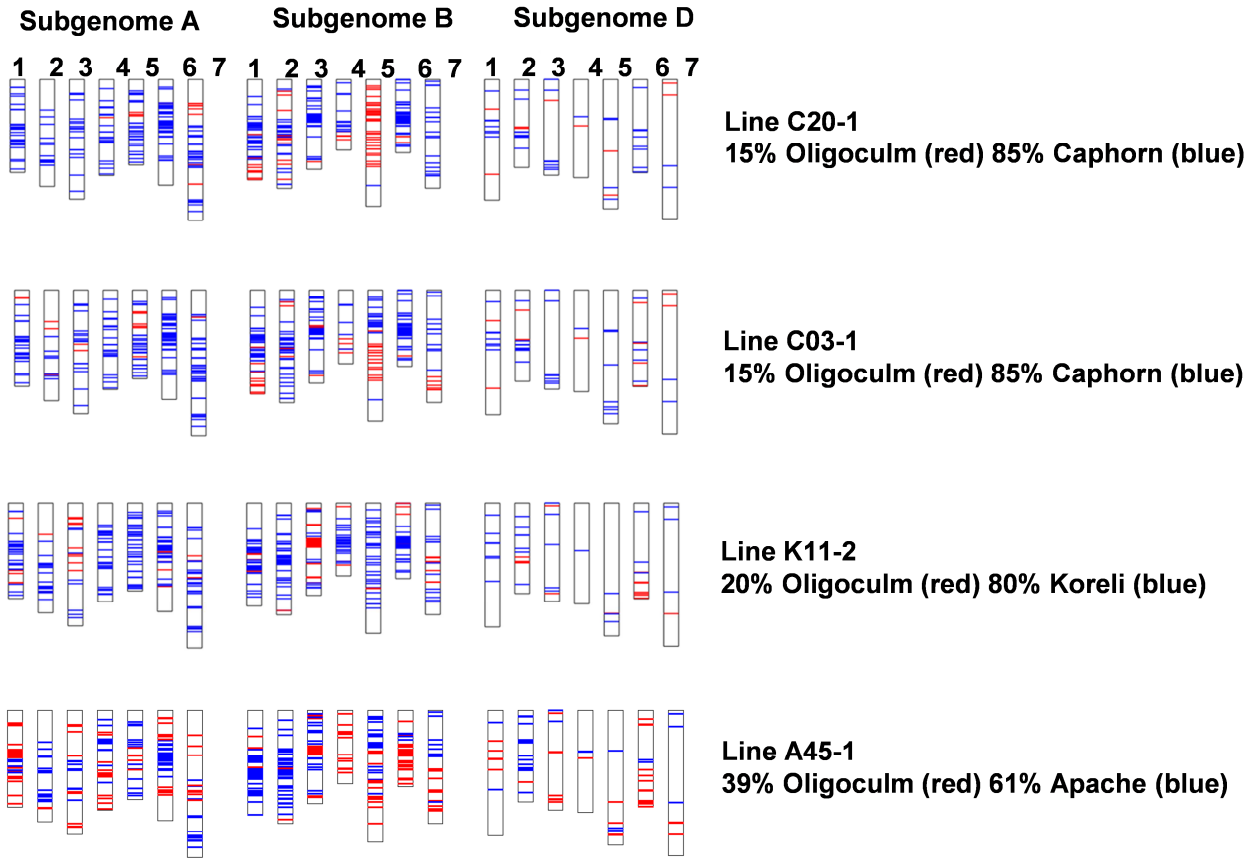
Supplementary Table 1: Molecular markers.

The table provides the list of molecular markers (name and primer sequences) used in the current analysis.

RJM246_F	TGTGCACATGTTCACAGCTC
RJM246_R	CCTAATTTGCGCCCACTCTC
RJM252_F	CTAGCGTGCCTTTTCCATCC
RJM252_R	AAGAGGAGGTGGAGGGGTAT
SSR01_F	CTCTTCACGGCGGAAGGT
SSR01_R	TTGGTCGGTAGGATTCAGCC
SSR29f	TTGATGTAACATGGCGAGCC
SSR29r	AACACGCCCAGGAAGATGAT
GWM136-L	GACAGCACCTTGCCCTTT
GWM136-R	CATCGGCAACATGCTCATC
cfa2153-f	TTGTGCATGATGGCTTCAAT
cfa2153-r	CCAATCCTAATGATCCGCTG
COS257-f	TCAAGCTTCTGCTTGAACACA
COS257-r	GGAGGGAAACAGCAGCAA
RJM.8F	TGGATACATGGTCAACAGTTTTT
RJM.8R	GGAAGAAGACGATGCACTCC
R23	GCTGACACGGGTTTTTAT
F24	CATTGCCAGCATACATTCTC

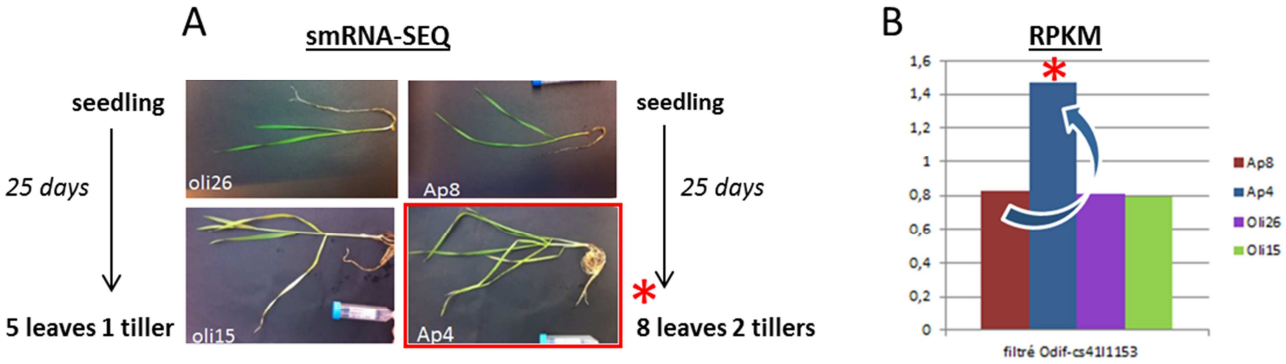
Supplementary Figure 2:

Oligoculm (TIN+) introgression in Koreli, Apache, Caphorn genotypes. Genotyping data from a 420K SNPs array unravelling the Oligoculm alleles (in red) and the parental alleles (in blue) for the four recombinant lines (C20-1, C03-1, K11-2, A45-1) at the TIN locus.



Supplementary Figure 3: miRNA1436 Expression analysis.

The expression was obtained by smallRNA sequencing on the recombinant lines Apache TIN- and Oligoculm TIN+. For each genotype, two plantlets (root-stem-leaf) were used at two developmental stages (delayed with 25 days). 4 contrasted samples were sequenced: Oligoculm seedling (oli26) and Oligoculm with 5 leaves and one tiller (oli15), the recombinant line Apache TIN- seedling (Ap8) and with 8 leaves and 2 tillers after 25 days (Ap4). (A) Pictures showing the plant material with oli26, oli15, Ap8, Ap4. (B) Gene (miRNA1436) expression (RPKM) on the 4 samples. The red star highlights the increased expression of miRNA1436 for the TIN-sample.

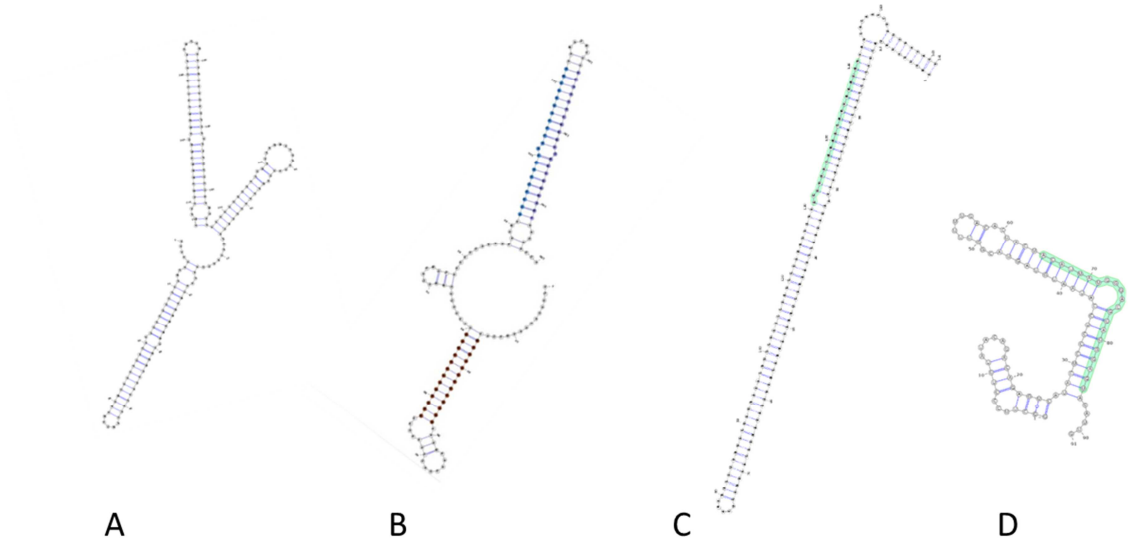


Supplementary Table 2: Molecular markers. miRNA1436 targets. From the psRNATarget (http://plantgrn.noble.org/psRNATarget/) using input:>mir1436ACTCCCTCCGTCCCATAATGT, preloaded transcript/genomic library for target search : Triticum aestivum(wheat), unigene, DFCI Gene Index (TAGI),version 12, released on 2010_04_18. The maximum expectation is the threshold of the score. A small RNA/target site pair will be discarded if its score is greater than the threshold. Target accessibility - maximum energy to unpair the target site (UPE) : The less energy means the more possibility that small RNA is able to contact (and cleave) target mRNA.

miRNA_Acc.	Target_Acc.	Expectation	UPE	miRNA_start	miRNA_end	Target_start	Target_end
mir1436	BE516875	0.0	22.904	1	21	179	199
mir1436	TC415196	0.0	13.479	1	21	680	700
mir1436	TC376396	0.0	18.462	1	21	700	720
mir1436	TC370353	0.0	19.095	1	21	726	746
mir1436	TC381627	0.0	17.111	1	21	894	914
mir1436	CJ881430	0.5	18.508	1	21	401	421
mir1436	TC408386	0.5	18.253	1	21	544	564
mir1436	CV775565	0.5	16.043	1	21	668	688
mir1436	CV772007	0.5	16.996	1	21	427	447
mir1436	CJ580639	0.5	19.15	1	21	9	29
mir1436	TC391933	0.5	23.486	1	21	635	655
mir1436	TC443783	0.5	17.188	1	21	1205	1225
mir1436	TC376770	0.5	15.365	1	21	1696	1716
mir1436	TC451725	1.0	20.751	1	21	534	554
mir1436	TC388133	1.0	19.599	1	21	317	337
mir1436	TC386636	1.5	17.951	1	21	839	859
mir1436	CA676402	1.5	18.807	1	21	57	77
mir1436	TC430344	1.0	18.58	1	21	298	318
mir1436	CD911932	1.5	22.642	1	21	425	445
mir1436	CJ699970	1.0	17.211	1	21	195	215
mir1436	TC383548	1.5	13.051	1	21	687	707
mir1436	CJ616995	1.0	15.151	1	21	11	31
mir1436	TC456462	1.5	21.366	1	21	175	195
mir1436	CO349548	1.5	23.954	1	21	199	219
mir1436	CA693005	1.5	9.602	1	21	31	51
mir1436	TC433863	1.5	13.177	1	21	444	464
mir1436	CN010452	1.5	18.047	1	21	554	574
mir1436	TC419136	1.5	17.942	1	21	624	644
mir1436	TC375086	2.0	17.154	1	21	625	645
mir1436	TC400854	2.0	16.5	1	21	636	656
mir1436	TC390151	1.5	12.631	1	21	856	876
mir1436	DR731606	1.0	22.738	1	20	238	257
mir1436	CK154549	1.0	21.735	1	20	254	273
mir1436	TC455761	2.0	12.035	1	20	4	23
mir1436	CA744134	2.0	17.042	1	20	4	23
mir1436	TC418296	2.0	13.612	1	21	325	345
mir1436	CK195702	2.0	16.073	1	21	68	88
mir1436	CJ906094	2.5	16.812	1	21	296	316
mir1436	CJ583468	2.0	19.689	1	21	46	66
mir1436	CJ951103	2.0	20.309	1	21	394	414
mir1436	TC390819	2.5	15.068	1	21	672	692
mir1436	TC426273	2.0	20.571	1	21	579	599
mir1436	TC412501	2.0	19.966	1	21	1178	1198
mir1436	TC461453	1.5	21.205	1	20	381	400
mir1436	BJ323295	1.5	6.975	1	20	48	67
mir1436	TC402807	1.5	15.363	1	20	752	771
mir1436	TC393171	1.5	15.337	1	20	932	951
mir1436	TC455466	2.0	15.113	1	21	241	261
mir1436	TC447101	2.0	20.342	1	21	232	252
mir1436	TC441175	3.0	20.216	1	21	751	771
mir1436	TC382242	2.0	16.581	1	21	763	783
mir1436	TC403387	2.0	23.835	1	21	1004	1024
mir1436	TC451364	2.0	19.196	1	21	1380	1400
mir1436	TC410264	2.5	16.955	1	21	1352	1372
mir1436	BE497599	2.0	19.934	1	20	73	92
mir1436	TC445508	2.0	13.042	1	20	4	23
mir1436	CJ868604	2.0	16.092	1	20	147	166
mir1436	CA608693	2.0	24.496	1	20	265	284
mir1436	GH722779	2.5	17.157	1	21	61	82

Supplementary Figure 4: miRNA1436 secondary structure.

Secondary structure predicted with the software STAR and visualized with VARNA (Darty et al., 2009). (a) ta-miR1436 from oligoculm TIN+ allele (b) ta-miR1436 chinese spring allele (c) osa-miR1436 (d) huv-miR1436



RESUME

Dans l'alimentation humaine, le blé joue un rôle capital du fait de sa valeur nutritive. Une hausse de la production de plus de 20 % sera nécessaire d'ici 2050 simplement pour garantir aux populations les standards actuels de consommation alimentaire. Prenant en compte les bouleversements climatiques créant des contraintes environnementales conséquentes, l'amélioration du rendement en blé sans perte de qualité devient un réel défi mondial. C'est dans ce contexte que s'inscrit ma thèse.

La génomique translationnelle est une approche intégrative qui fait le lien entre Recherche Fondamentale et Appliquée, où les espèces modèles jouent le rôle de pivot pour étudier les espèces d'intérêt agronomique. J'ai mis en œuvre cette approche de recherche translationnelle pour étudier finement l'histoire évolutive, l'organisation et la régulation du génome du blé. Le blé est une espèce polyploïde qui a subi des duplications chromosomiques récentes (500 000 et 10 000 ans) et anciennes (<90 millions d'années). Mes travaux ont consisté à utiliser les espèces de céréales apparentées pour étudier l'impact de ces duplications sur la plasticité structurale et expressionnelle des copies de gènes dupliqués du blé moderne.

Mes travaux ont montré que la polyploïdie chez le blé est suivie d'une diploïdisation. Cette diploïdisation est en cours chez le blé moderne ; elle consiste en l'accumulation de mutations, de perte de gènes ou de modification de l'expression des gènes dupliqués. Cette diploïdisation est non aléatoire ; elle génère des blocs chromosomiques dominants à forte stabilité et d'autres plus sensibles, à forte plasticité. Au travers de l'analyse du génome du blé, la polyploïdie apparaît comme une force majeure de l'évolution, voire de l'adaptation, en permettant la spécialisation structurale et fonctionnelle des gènes surnuméraires. Cette asymétrie de plasticité structurale et expressionnelle post-polyploïdie entraine *in fine* la diploïdisation des phénotypes. Mes travaux de thèse l'illustre au travers de l'analyse des bases génétiques de l'inhibition du tallage, contrôlée par une insertion de 109bp codant pour un microRNA porté uniquement par la région chromosomique 1A, dite sensible.

Mes travaux montrent une quasi-complète diploïdisation structurale, expressionnelle et phénotypique du blé tendre moderne ouvrant la question d'une re-définition du concept « d'espèces polyploïdes » au regard des analyses génomiques qui peuvent être conduites aujourd'hui, comme cette thèse en est une illustration.

ABSTRACT

Wheat plays a key role in Human food due to its nutritional value. Wheat production needs to be increased by more than 20% by 2050 to guarantee current human consumption standards. Taking into account climatic changes with high level of environmental constraints, yield improvement without quality loss became a big challenge. This consists in the economical and societal context of the current doctoral thesis.

The integrative translational genomic approach consists in transferring fundamental knowledge gained from model species to applied practices for breeding in crops. This strategy was used here to study the evolutionary history, the organization and the regulation of the modern bread wheat genome. Modern wheat is a polypoid species deriving from two hybridization events between diploid progenitors 500 000 and 10 000 years ago, as well as a more ancient that dated back to more than 90 million years ago. The current research consisted in using cereal species closely related to wheat to study the impact of these duplications on the structural and expression plasticity of duplicated genes in wheat.

My results established that the diploidization process is in progress in wheat after the successive rounds of polyploidization events. This diploidization consists in the accumulation of mutations, gene loss or expression modification between duplicated genes. This diploidization is nonrandom at the genome level; generating dominant chromosomic regions with high stability in contrast to others regions more sensitive with high plasticity. Based on such wheat genome evolutionary analysis, polyploidy appears as a major evolutionary force driving plant adaptation through structural and expressional specialization of duplicated genes.

Such post-polyploidy genomic asymmetry drives finally the phenotype diploidization as illustrated in the current research with the study of genetic basis of the tiller inhibition Trait. This trait seems to be driven by a 109 pb insertion coding for a microRNA located solely on the chromosome 1A, known as a sensitive genomic fraction.

The current research established that the modern bread wheat has been quasi-entirely diploidized at the structural, expressional and phenotypic levels, now requiring a new definition of the polypoid concept in line with current genomic investigations, as illustrated in the current thesis.