Expression et régulation épigénétique des gènes homéologues chez le blé tendre

Caroline Juery

Ecole doctorale des Sciences de la Vie, Santé, Agronomie & Environnement

24 novembre 2020

Thèse de doctorat présentée à l'Université Clermont Auvergne pour l'obtention du grade de

Docteur d'Université spécialité Biologie Végétale

Soutenue par

Caroline Juéry

Expression et régulation épigénétique des gènes homéologues chez le blé tendre.

Membres du jury :

Karine Alix, Professeur, HDR, AgroParisTech, – Rapportrice
Marie-Angèle Grandbastien, DR, HDR, INRAE– Rapportrice
Malika Ainouche, Professeure, HDR, Université de Rennes – Rapportrice
Christophe Tatout, Professeur, HDR, Université Clermont Auvergne - Examinateur
Etienne Paux, DR, HDR, UMR GDEC Clermont-Ferrand (Directeur de thèse)
Frédéric Choulet, IR, UMR GDEC Clermont-Ferrand (Co-encadrant de thèse)

UMR 1095 INRA-UCA, unité GDEC « Génétique, Diversité et Ecophysiologie des Céréales »
5 chemin de Beaulieu, 63000 Clermont-Ferrand

« L'écriture est une aventure. Au début c'est un jeu, puis c'est une amante, ensuite c'est un maître et ça devient un tiran ».
Winston Churchill

« On peut tuer un homme, mais pas ses idées ».
Thomas Sankara

"La poésie a, comme la vie, l'excuse de ne rien prouver."
Emil Michel Cioran

« La science consiste à oublier ce que l'on croit savoir ; la sagesse à ne pas s'en soucier »
Charles Nodier

Remerciements

Mercis pour le projet

Tout d'abord merci à **Etienne** et **Fred**, mes encadrants pour ce projet de thèse, que j'avais rencontré en 2015 pour discuter d'épigénomique chez le blé tendre. Merci à vous deux de m'avoir accueilli dans votre équipe SEVEN et de m'avoir aidé jusqu'au bout dans ce projet. Ma gratitude va de pair avec des excuses, pour ne pas avoir été à la hauteur de vos espérances et pour les difficultés que j'ai pu vous causer. Merci pour ces moments de discussion scientifique et sur l'état de la recherche en général. Merci **Etienne** pour ces discussions pour tenter de recadrer mon esprit fougueux et éviter de partir dans des hypothèses trop bancales. Je me souviens de ce premier congrès à Paris où avait fini par discuter un peu littérature au restaurant, prémonitoire ? Merci **Fred**, j'ai vraiment apprécié ces discussions de fin de journées, 17h30-18heures, tu poussais la porte du bureau pour savoir comment ça allait et on finissait par discuter évolution, recombinaison et statut du chercheur. Des moments simples et motivants.

Merci **Thierry** et **Pierre** pour votre soutien et votre compréhension, surtout en cette fin difficile. J'ai beaucoup apprécié les discussions du confinement sur l'évolution des espèces et les hypothèses un peu marginales formulées par certains chercheurs.

Merci **Plateforme Gentyane**, merci **Charles**, **Véro**, **Mickaëla**, **Elodie, Carole, Anthony** pour ces moments ou j'essayais de faire du ChIP-seq. J'ai squatté votre Fragment Analyser un peu trop souvent ;-). En tout cas merci pour vos conseils et votre aide et aussi pour la bonne humeur au PAG !

Merci **David Latrasse** et **Moussa Benhamed** pour votre formation sur le ChIP-seq au sein de votre laboratoire. Merci **David** pour tous tes conseils à distance.

Merci **Région Auvergne**, désormais fusionnée avec ta sœur Rhône-Alpes, pour les financements « sans-conditions » ni obligation de résultats et surtout pour tes **volcans** et tes **lacs** qui permettent de se ressourcer pendant les moments d'épuisement. Merci l'**INRA** et le **GDEC**, établissement fer de lance de la génétique sur le blé, où je me suis souvent senti comme un éléphant dans un jeu de quille ! Merci pour votre accueil.

Mercis pour la bio-info et R

Merci **Romain** pour ton aide si précieuse dans mes débuts en bio-informatique. Bash, sed, awk mais aussi et surtout ces fameuses tables de hachages que je n'ai pas vraiment appris à dompter. Sans toi je n'aurais rien fait puisque ce sont des pages et des pages de tes lignes de commandes qui sont enregistrées dans le fichier « commandes_trop_cool.sh » et avec lesquelles j'ai travaillé de longues

heures. Et toutes mes excuses pour ces moments de folie pendant lesquels je ne pouvais pas m'empêcher de me prendre pour Freddy Mercury ;-)

Merci **Hélène**, toi qui m'as apporté ton soutien info/bioinfo pour installer des packages, mieux comprendre les données « omics » ; je me souviendrais toujours de ce premier script python que tu m'avais rédigé en une heure ou deux, je me suis dit « oulalala… mazette… comment elle fait ça ? ».

Merci **Lorenzo** pour ton aide sur la partie épigénétique, mapping des reads et peak-calling de données ChIP-seq, méta-genes profiles. Je n'ai pas pu m'en servir directement pour mes résultats mais merci pour le temps que tu m'as accordé, à distance, pour m'expliquer et me rassurer vis-à-vis de mes lacunes.

Merci **Jo** pour tes compétences Figures R que tu as partagées avec moi, ainsi que ton soutien technique pour les premières manips ChIP-seq. Merci **Nathan** pour ton aide pour « A la recherche des homéologues chez les espèces diploïdes », digne d'un Indiana Johnes !!

Et merci **Fred** pour avoir accordé de ton temps en dernière année de thèse et proposé une formation aux néophytes en bash. De petits suppléments de commandes qui m'ont permis de sortir des sables mouvants. Merci pour toutes les fois où tu es venu voir si je m'en sortais.

Merci **Sarah** et merci **Agathe** pour votre aide précieuse sur les scripts R. Là aussi pas mal de fois bloquée, vous avez su m'écouter et me proposer des solutions sophistiquées pour répondre à mes questions. MERCI. Merci **Cécile**, **Kévin**, merci **Nathan**, merci **Marion** pour votre bonne humeur, discussions science et conseils ☺ Merci à **toutes les personnes que j'ai croisées dans les couloirs**, les matins, les soirs, les we, la journée qui m'ont demandé comment j'allais et si j'avançais. Merci pour vos attentions bienveillantes, c'est si important (**Isa**, **Bouzid**, **Pascal**, **Cybille**, **Ludo**, **Annaïg**, **Magalie**…).

Merci à mes **correcteurs orthographiques** qui sont des humains qui ont appris et retenu leurs règles de grammaire et d'orthographe et non des logiciels : **Mélisande, Adeline, Maman, Guitou, Claudine Marion, Nadège, Tetelle.**

Mercis à l'amitié, à la vie, à vous

Merci **Fatiha.** C'est simple, sans toi, je serais restée au bord de la route. Tu m'as donné de ta force, de ton courage, tu as été les bras qui étaient là pour me retenir quand je trébuchais dans ce parcours d'obstacles. Combien de fois as-tu séché mes larmes ? Tu m'as fait découvrir ton pays, tes racines et tu m'a emmenée jusqu'au bout de l'aventure. Tu es devenue une amie, de celles qui sont là, qui restent, pour toujours. Merci et longue vie à cette amitié ☺.

Merci **K** ! Benoît Kéraval, premier correcteur de presque tous les documents que j'ai pu écrire… Comment je t'ai emm***é ! Mais qu'est-ce que c'était cool ces discussions scientifiques ! On a appris à débattre ensemble, maintenant on n'a plus envie de se taper quand on part sur les chemins de la science ☺ Merci pour ton soutien infaillible ! Longue vie à cette amitié aussi ☺

Marion, Julie, une amitié de 11 ans, toujours aussi forte pour soutenir la plus fragile de nous trois, vous me connaissez, ça me met les larmes aux yeux ☺ ! Vous avez toujours été là, à 5h du matin comme à minuit, mots rassurants, bienveillance immense, sans relâche. Merci pour votre témérité dans cette amitié.

Robin Michard, je te décerne une maîtrise en psychologie spécialité doctorants en détresse et problèmes sentimentaux. Ton calme en toutes circonstances et tes mots rassurants m'ont fait beaucoup de bien. Je ne sais pas comment tu fais, tu m'impressionnes.

Merci **Tételle**, Reine des soirées jeux et des randos qui font du bien. Merci pour proposer toujours des moments de retrouvailles simples et chaleureux ! Merci de m'avoir fait découvrir ce puy de Côme que l'on aime tant.

Merci la **bande à doudous** ! **Doudou** (Fred), **Doudoue** (Kristell), **Doudounette** (Charlotte), **Bill** alias **Kéraval**, **Tételle** épouse Hubert, la **Mère Hubert** épouse Kreen, **Vinc'ouille** la fripouille, **Quot la Tourmente** deuxième fripouille, **Pépé**, **Caro**, **Mathoole**, **Barbier**. Merci pour ces repas, mariages, bébés, soirées et aventures de pirates et pour toutes les fois où vous m'avez dit que ça allait le faire ! Un jour Jacklyn Sparrow vous emmènera sur son bateau, parole de pirate ☺

Merci **Coloc N°1** : **Benoit (**Sir Kéraval**)**, **Claire (**Dame Lefèvre**)**, **Guigui (**Lord Rivoalland**,** désormais vrai Lord**)**. Cette année et demie passée à vos côtés a été riche en couleurs avec cette bonne ambiance de colok de zinzins, projets construction, apéros cacahuètes, virée au porge, virée ski en savoie, c'était trop cool ! Vous avez été en première ligne des débuts de ce doctorat. C'était vraiment chouette de retrouver la bonne ambiance à la maison les soirs d'infortune ! Zinzins for ever ! Merci **JuJu** pour ta crête punk, les réparations vélo, les montages de barbuc, les conseils d'escalade.

Merci **chemins d'errance** et merci **Loïc**, **Benoît** et les **flancs de la faille de Limagne** pour m'avoir hébergée les soirs d'été 2019 ou je suis devenue SDF.

Merci **Coloc N°2** : **Mélisande**, **Benjamin**, **Adeline**. Merci Mel pour ce projet coloc, on se comprend, on est tellement sur la même longueur d'ondes, thésardes hypersensibles, je te considère comme ma deuxième sœur ☺. Merci pour ces débats sans fin, ces veillées ou on refait le monde à coup d'infusions et de questions. Où se trouve l'alternative à cette société malade ? Et les personnes à handicaps mental, comment on s'en occupe ? Et l'économie de demain c'est quoi ? C'est quoi les propriétés du nombre Pi Benjamin ? Mélisande, c'est quoi le problème majeur qui freine le passage de

l'agriculture productiviste à l'agriculture intégrée dans les écosystèmes ? Merci Adeline pour tes petits mots doux et les poivrons rouges ;-) Vélorution, la voiture c'est la pollution !!!

Un grand merci à vous **Claudine** et **Dominique** pour m'avoir offert gîte et couvert maintes fois et pour tous ces moments d'échanges profonds sur la pédagogie, le sens du travail pour la communauté, les institutions, leurs devoirs, leurs dérives. Merci de m'accompagner de votre expérience et de votre recul sur les choses pour beaucoup de mes projets, parfois pas tous très raisonnables ;-).

Merci à la **familia**, **Papa**, **Maman**, **Marc**, **Audrey**, **Guillaume**, **Marie**, **Etienne** pour votre soutien, les repas en famille, les travaux à la ferme qui font du bien pour se changer les idées et merci aux petits nouveaux **Margaux**, **Tom**, **Emma** pour leur énergie et leur insouciance !

Merci **Sarah** pour la motivation piscine, pour toutes les fois où tu m'as rassurée et proposé des petits repas ou soirées.

Merci **Loïc** pour la phénoménologie et la photo argentique. Et merci le **Labomatik** !

Merci à **Alexandra Assanovna Elbakyan**, fondatrice de Sci-hub. Sans toi, aucun manuscrit possible !

Merci à tous ceux que j'ai oublié, j'ai pas fait exprès… mais je pense à vous aussi qui avaient été là (**Josie**, **Marielle**, **Mathilde Lebot, Lucie, Horacio, Béné, Nico, Noémie, Audrey…**).

Mercis aux artistes et philosophes

Merci à tous les artistes qui m'ont accompagnés : **Freddy** et le groupe **Queen** pour ton énergie, merci **Patti** pour ta poésie, Merci **Lou** et les **Velvets** pour votre rock grunge garage, pour ce solo de guitare sur Sweet Jane, merci **Jean-Louis** pour ta douceur, merci **Alain** pour ton élégance, merci **Nina**, merci **Aretha** pour votre force, merci **Sona Jobarteh**, merci les Ritas, merci les Rolling Stones, merci Jimmy, merci Janis, merci Jim Morison, merci Manu, merci Soldat Louis, merci Didier et les Wampas, je vous aime, merci Ska-p, merci Etienne, merci Led Zep', merci les Pink Floyd, merci The Animals, merci Clara, merci Céline, merci Curt, merci **David Bowie** pour ta folie, merci Bob, merci Oasis, merci Bernard, merci Georges, merci **Nino**, merci Bourvil, merci Henri, merci **Jean-Jacques**, merci Serge, merci **Eddy**, merci les **Pink Martini**, merci Gaëtan, merci **Jacques**, merci la Rue ket' (y'a parfois des cigales dans la fourmilière et on peut rien y faire), merci les Soufris (chienne de vie chienne de vie chienne de vie…), merci Florent, merci Louis, Merci M, merci Les Innocents, merci Blankass, merci Matmatah, merci Francis, merci Alanis, merci Beyonce, merci la **Compagnie Créole** ; merci à vous philosophes : **F.Nietzsche**, merci **B. Spinoza**, merci **Aristote**, merci **Socrate**, merci **Platon**, merci **Gaston** Bachelard, merci Arété de Cyrène première femme reconnue comme philosophe… et tant d'autres. Vous m'avez accompagné sans le savoir sur ce chemin escarpé, c'était chouette avec vous.

Résumé

De nombreuses espèces de plantes sont polyploïdes, c'est-à-dire qu'elles possèdent plusieurs sous-génomes au sein du noyau de leurs cellules. La polyploïdie s'accompagne d'une redondance génétique qui offre un potentiel d'innovations évolutives important par un relâchement de la pression de sélection autorisant sous-fonctionnalisation, néo-fonctionnalisation, perte de gènes. Le blé tendre est une espèce polyploïde récente, apparue suite à deux hybridations interspécifiques (800 000 et 10 000 ans). Il possède un génome hexaploïde composé de trois sous-génomes : AABBDD et théoriquement, il possède trois copies homéologues de chaque gène (1A:1B:1D). Cependant, les analyses génomiques ont révélées que la moitié des séquences codantes présentaient un nombre de copie de type NA:NB:ND. Comment évolue cette redondance génétique après la polyploïdisation chez le blé tendre? Peut-on observer des différences d'expression des copies de gènes témoignant d'une évolution fonctionnelle pour cette espèce formée très récement? Quels sont les mécanismes sous-jacents ?

L'objectif de cette thèse a été d'analyser les expressions relatives des copies de gènes homéologues pour des groupes présentant trois (1 :1 :1, triades), deux (0:1:1, 1:0:1 ou 1:1:0, dyades) ou quatre copies (2:1:1, 1:2:1 ou 1:1:2, tétrades). Nous avons également relié les résultats aux caractéristiques structurales (position génomique), évolutives (présence ou absence des copies chez les espèces ancêtres) et épigénétiques (marques histones) des gènes pour répondre aux questions de recherche. Nous avons utilisé les données de RNA-seq et de ChIP-seq mises à disposition lors de la publication dela séquence génomique de référence du blé tendre (IWGSC 2018).

Nous avons mis en évidence que les 51,1% de gènes en triades présentent en majorité (81%) une expression équilibrée sur l'ensemble des tissus et au cours du développement (expression élevée et constitutive). Ces gènes sont majoritairement associés la marque épigénétique d'activation de l'expression : H3K9ac. *A contrario*, les gènes en dyades (11,7% des gènes) et en tétrades (2,8% des gènes) présentent plus fréquemment des biais d'expression (36% et 75,4% respectivement). Ces gènes sont plus associés à la marque épigénétique liée à la répression ciblée et transitoire des séquences (H3K27me3). En revanche, aucune dominance d'expression n'a été décelée à l'échelle du génome entier. Ceci met en évidence de potentielles sous-fonctionnalisations des gènes, plus fréquentes pour des gènes différents des triades, présents dans les régions distales des chromosomes. Même si les biais d'expression correspondent à des différences déjà existantes chez les espèces ancêtres, nous avons cependant distingué des traits d'expression correspondant aux différentes étapes de l'histoire évolutive du blé : les copies du sous-génome D sont moins réprimées et moins associées à la marque H3K27me3 ; les biais d'expression entre les copies AABB sont plus prononcés. Ainsi, la coévolution des deux sous-génomes AABB pendant 800 000 ans est décelable alors que le sous-génome D semble encore s'exprimer de façon autonome.

Ces résultats suggèrent que ce génome comprend des gènes très contraints évolutivement qui constitueraient le « core » génome de l'espèce avec des fonctions de bases conservées (gènes en triades) et des gènes présentant des variations du nombre de copies, des régulations différentielles et des fonctions spécifiques témoignant de possibles innovations évolutives, appartenant probablement au génome dit « dispensable » (dyades et tétrades).

Mots clefs : polyploïdisation – *Triticum aestivum* – biais d'expression homéologues – épigénétique - évolution

Abstract

Within the plant kingdom, a lot of species are polyploids, meaning that they present two or more sub-genomes in the nucleus of their cells. Polyploidy confers genetic redundancy that offers a high potential of innovations and adaptations by relaxing natural selection on genic sequences. This allows faster sub and neo-functionalization of genes but also a loss of sequences that might be stochastic or not between the sub-genomes. Bread wheat is a recent polyploidy species that derived from two interspecific hybridizations that occurred 800 000 and 10 000 years ago. The genome of this species contains three sub-genomes: AABBDD and in theory three copies of each gene (1A:1B:1D). However, genomic analysis of the genome sequences reveals that half of the genes present copy number variations (NA:NB:ND). Within this scientific context, we wanted to answer questions such as: How this genetic redundancy evaluates after the polyploïdisation process? Is-it possible to observe differences in terms of gene expression that could correspond to functional evolution for this recently formed species? Which mechanisms could explain those processes?

The objective of this PhD was to analyses relative expressions of homoeologous genes of bread wheat for groups presenting one copy on each sub-genomes (1 :1 :1, triades) and groups presenting a copy number variation with a loss (0:1:1, 1:0:1 ou 1:1:0), dyads or a duplication (2:1:1, 1:2:1 ou 1:1:2, tetrads) of sequences. We linked this analysis to genomic characteristics such as chromosome structure (genomic position of genes for exemple), evolution (presence or absence of lost and duplicated copies within diploid genomes of the progenitor species) and epigenetics (histone modifications). We used RNA-seq and ChIP-seq data released at the same time as the publication of the genomic reference sequence of bread wheat (IWGSC 2018).

We highlight that the 51,1% of triads genes present mostly (81%) a balanced expression across the 15 tissues and developmental stages analyzed (high and constitutive expression) Those genes are mainly associated with the H3K9ac histone mark that is linked to an active transcription of genes. At the opposite, dyad genes (11,7% of High Confidence wheat genes) and tetrad genes (2,8%) present more frequently unbalanced expression patterns (36% and 74,5% respectively). Those genes are more associated with the histone mark H3K27me3 defining facultative heterochromatin and that target genes with transient expression. No dominance of one sub-genome on the others was discovered at the whole genome scale but rather stochastic suppression of genes copies. These results reveals potential sub-functionalization of genes, more frequent for copies present I the distal regions of chromosomes and associated with the epigenetic mark H3K27me3. Even if the homoeolog expression bias mostly corresponds to already existing divergence between diploid progenitor species, we nevertheless observe expression bias corresponding to the different step of bread wheat evolutive history: copies from sub-genome D are less repressed than the A or B copies; expression bias between AABB copies are more pronounced. In that respect the co-evolution of the two sub-genomes AABB during 800 000 years are traceable while D sub-genome seems to still present a nearly autonomous expression

Combined together, these results suggest that wheat genome contains genes evolutionary constraints that correspond to a “core” genome of the species with basic conserved function (triad genes) and genes that present variation of the number of gene copies with differential regulations and specific functions that correspond to “dispensable” genes (dyads and tetrads).

Key words : polyploïdization – *Triticum aestivum* – homeolog expression bias – Epigenetic – Genome évolution .

Liste des abréviations

ADN : Acide DésoxyRibonucléique

ChIP-seq : Chromatine Immunoprecipitation – sequencing

CS : Chromatin State

GO : Gene Ontology

HC : Hight Confidence

HomoeoCNV : Homoeologous Copy Number Variations

IWGSC : International Wheat Genome Sequence Consortium

LC : Low Confidence

Mb : Mégabase

MYA : Million Years Ago

PAV : Presence Abscence Variation

PPD : Post-Polyploidization Diploidization

RdDM : RNA directed DNA methylation

RNA-seq : RiboNucleic Acid - sequencing

TE et **ET**: Transposable Elements et Eléments Transposables

TPM : Transcript Per Million

UTR : UnTranslated Region

WGD : Whole Genome Duplication

Liste des Figures

Liste des tableaux

Table des matières

Introduction générale

Depuis plus de 3,5 milliards d'années, la matière vivante s'est constituée sur la planète Terre et ne cesse d'évoluer, prenant diverses formes toujours plus surprenantes et sophistiquées pour nous humains, observateurs de ce processus. Ceci est d'autant plus fascinant chez les plantes qui sont des organismes fixés dans leur milieu, devant ainsi affronter les variations environnementales parfois extrêmes. Cette évolution des formes de vie s'explique par l'apparition de variations liées à la plasticité phénotypique et à la diversité génétique se traduisant par des innovations adaptatives sélectionnées au cours de l'évolution des populations d'individus de l'espèce. La plasticité phénotypique adaptative peut être en partie liée aux changements épigénétiques modifiant l'expression des gènes en réaction à des stimuli environnementaux. La diversité génétique est quant à elle encodée dans le génome et s'explique par l'apparition de mutations aléatoires des séquences codantes ou des séquences régulatrices qui peuvent être sélectionnées au cours de l'évolution. Cela peut ainsi favoriser l'apparition de protéines exprimées différement ou aux fonctions nouvelles. En plus de ces deux grands principes, un processus biologique a été mis en évidence comme ayant grandement influencé l'évolution des génomes, en particulier celle des génomes végétaux : la duplication de génome encore appelé phénomène de polyploïdisation.

Découverte dans les années 1910 par l'analyse cytogénétique de noyaux de cellules grâce au développement de la microscopie, la polyploïdie correspond à l'état génomique d'une espèce dans lequel plus de deux jeux de chromosomes sont retrouvés dans le noyau de ses cellules somatiques (Ramsey et Ramsey 2014). Une espèce diploïde comprend les n chromosomes du génome paternel appariés à n' chromosomes maternels donnant nn' paires de chromosomes. Pour les espèces polyploïdes ce sont nn' et mm' (voire plus) paires de chromosomes qui vont composer le génome, n et m représentant deux jeux de chromosomes différents appelés sous-génomes. Le génome comprendra de 2, 3, N sous-génomes (espèces tétraploïdes, hexaploïdes etc. respectivement composées de deux, trois, N jeux de chromosomes dits homéologues). Grâce au séquençage de milliers de génomes, nous savons maintenant que des évènements de polyploïdisation ont eu lieu chez tous les eucaryotes et que ce processus est particulièrement réccurent chez les végétaux (Van de Peer *et al.* 2017). La polyploïdisation semble même être un processus majeur de l'évolution des espèces végétales car cela permet d'apporter de façon soudaine une grande redondance génétique favorisant l'évolution fonctionnelle des séquences par relâchement de la pression de sélection (Mable *et al*. 2018). En effet, lorsqu'un génome présente plusieurs copies d'un même gène, l'une d'elle peut être affectée par un changement de régulation épigénétique ou une mutation dans sa séquence. Ces changements ne sont pas forcément contre-sélectionnés puisque l'une des copies peut assurer la fonction initiale. Les recherches effectuées dans ce domaine visent entre-outre à répondre à la question : « Quels processus permettent de créer de la diversité à partir de la redondance génétique ? ».

L'épigénétique est une discipline qui étudie l'ensemble des processus expliquant les variations de l'expression des gènes, héritables au cours des mitoses et/ou au cours des générations (Felsenfeld 2014). Ces processus correspondent pour l'essentiel à des modifications de l'accessibilité de la chromatine à la machinerie de transcription. Ils sont principalement régulés par des modifications biochimiques de la molécule d'ADN (densité de méthylation de l'ADN) et des protéines histones formant les nucléosomes qui permettent de condenser l'ADN dans le noyau (modifications post-traductionnelles des histones). La compréhension de ces processus sont cruciaux pour (i) mieux comprendre la plasticité phénotypique des espèces, (ii) déterminer le nombre de générations sur lesquelles ils peuvent être transmis, (iii) savoir s'ils sont sélectionnés au cours de l'évolution et peuvent directement participer à l'évolution des formes de vies. En particulier, participent-ils de l'évolution de la redondance génétique au sein d'un génome polyploïde ?

Comme beaucoup d'autres espèces d'intérêt agronomique, le blé tendre (*Triticum aestivum*) est une espèce polyploïde. Il est le produit de deux hybridations interspécifiques successives, ayant eu lieu il y a 800 000 et 10 000 ans environ qui ont conduit à la formation d'un génome qui comprend 3 sous-génomes A, B et D. En plus de cette redondance liée à la polyploïdisation, les espèces du genre *Triticum* présentent un fort taux de séquences dupliquées en comparaison avec d'autres céréales de la famille des poaceae (Sorgho, Brachypodium, Riz) (Glover *et al.* 2015). Ainsi, cette espèce qui présente un génome très redondant génétiquement constitue un modèle de choix pour étudier l'évolution des génomes et des séquences codantes.

Dans ce contexte, trois axes de recherche ont guidé ce doctorat. Le premier visait à estimer l'impact de la redondance génique sur l'expression des copies des gènes homéologues au sein du génome hexaploïde du blé tendre. Le second était de déterminer le rôle potentiel des variations épigénétiques sur la diversification de l'expression des gènes homéologues. Le troisième visait à produire des données de ChIP-seq (données omiques sur les marques épigénétiques) sur plusieurs tissus pour la marque histone H3K27me3 chez le blé tendre afin d'étudier plus précisément le rôle de cette dernière dans la spécificité d'expression des gènes homéologues. Afin de répondre à ces objectifs, ce manuscrit est organisé autour de 4 chapitres. Dans le premier chapitre est proposée une synthèse bibliographique présentant la polyploïdie, son impact sur le fonctionnement des génomes, le rôle potentiel de l'épigénétique et les caractéristiques génomiques de l'espèce d'intérêt (*Triticum aestivum*). Les deux chapitres suivants (présentés sous forme d'articles) présenteront les travaux conduits durant cette thèse portant sur (i) la caractérisation de l'expression des gènes homéologues présentant une seule copie sur chaque sous-génomes (gènes dits en triades) et (ii) la caractérisation de l'expression de copies de gènes homéologues présentant une absence de copie (dyades) ou une copie supplémentaire (tetrades) sur l'un des sous-génomes. Le chapitre III décrit les résultats d'expérimentation sur la technologie de ChIP-seq chez le blé tendre. Enfin, le chapitre 4 clôturera ce manuscrit avec une discussion de l'ensemble des résultats et des perspectives d'études.

CHAPITRE I

Synthèse bibliographique

I. Qu’est-ce que la polyploïdie ?

I.1 Définition et fréquence

La ploïdie est une notion qui caractérise le nombre de jeux de chromosomes au sein du noyau d’une cellule. La plupart des animaux sont diploïdes, avec des paires de chromosomes homologues, un paternel et un maternel, constituant un seul génome (AA’). On parle de polyploïdie lorsque le génome d’une espèce comporte plus de deux jeux de chromosomes dans les cellules somatiques, consittuant des sous-génomes (AA’ et BB’ par exemple). On parlera de polyploïdisation ou de doublement intégral du génome (Whole Genome Duplication en anglais ou WGD) pour signifier le processus conduisant à la formation d’une espèce polyploïde. Grâce au séquençage d’un nombre croissant de génomes, les analyses phylogénétiques ont révélé des évènements ancestraux de duplications de génomes et des espèces polyploïdes ont été identifiées chez tout type d’eucaryotes : les insectes (Li *et al.* 2018), les amphibiens (Schmid *et al*. 2015), les poissons (Zhou et Gui, 2017), quelques mammifères (Acharya and Ghosh, 2016), mais également chez les champignons (Albertin et Marullo 2012), chez les *Archae* (Markov et Kaznacheev 2016).

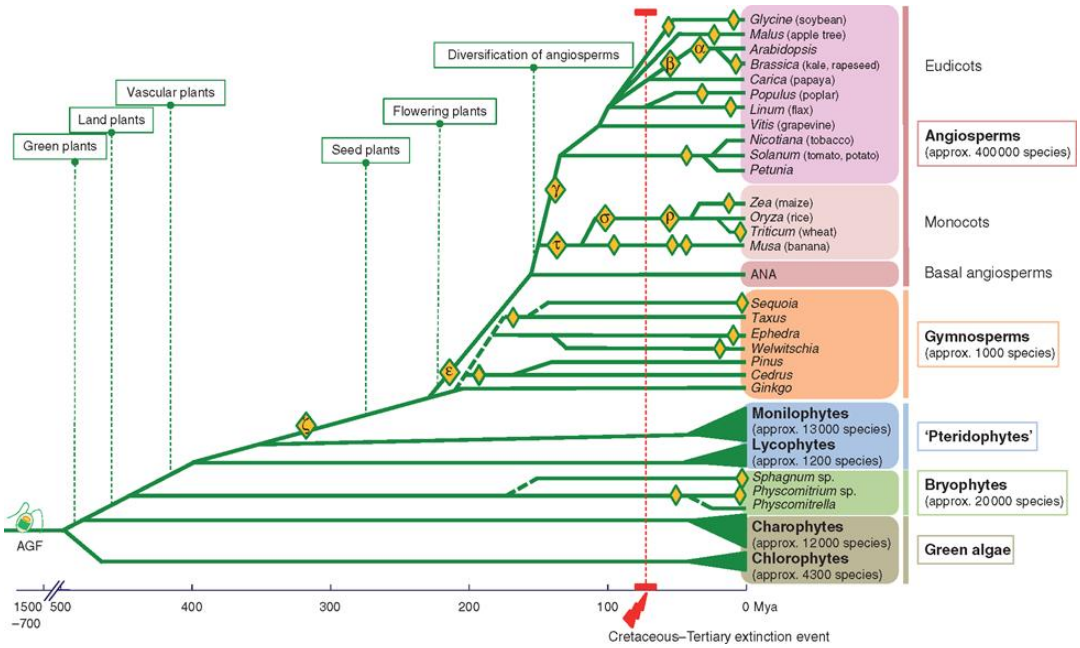


Figure 1. Phylogénie simplifiée des plantes de la lignée verte et évènements de polyploïdisation (WGD).

Losanges jaunes : évènements de duplication ou triplication. Il est possible de détecter la rapide radiation des angiospermes à partir des Mesangiospermae vers 139–156 Mya (Moore et al.2007, Bell et al.2010) avec un burst d’évènements de polyploïdisation dans la période du Crétacé après 125 Mya (âge du premier macrofossile d’angiosperme retrouvé ; Cascales-Miñana et al.2016). Les branches de l’arbre en pointillés représentent les parties d’arbres phylogénétiques encore imprécises en termes de temps ou d’apparentement entre espèces. Alix et al. 2017

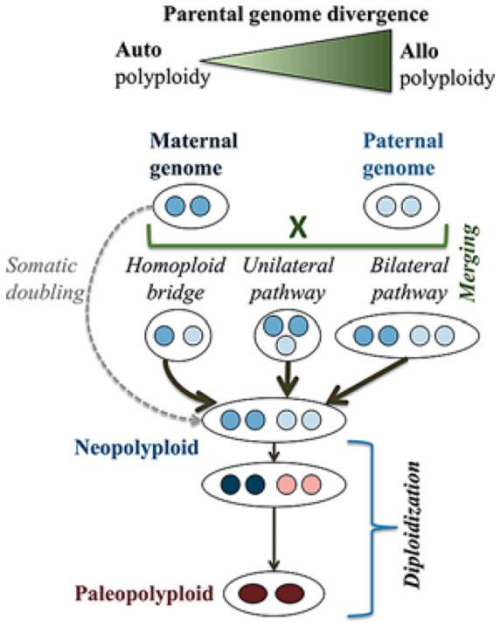
Les rares exemples retrouvés chez les métazoaires indiquent que le phénomène de polyploïdisation dans cet embranchement est plutôt contre-sélectionné. Pour l’expliquer, Wertheim *et al*. 2013 citent plusieurs études (Muller 1925, Mable 2004 par exemple) démontrant que le développement embryonnaire chez les animaux est très sensible aux variations de doses pour l’expression des gènes,

notamment pour ceux liés au déterminisme sexuel, ce qui expliquerait la faible fréquence d'animaux polyploïdes.

En revanche, la polyploïdie est particulièrement fréquente chez les plantes. L'état polyploïde des génomes est retrouvé chez un large spectre d'angiospermes, mais aussi chez des gymnospermes (Li *et al*., 2015), les fougères (Schneider *et al*., 2017) et les diatomées (Parks *et al*., 2018). Les travaux de phylogénétique analysant le partage de gènes dupliqués entre taxons végétaux ont permis d'identifier et de dater un premier événement ancestral de polyploïdisation à 320-350 millions d'années (Ma), antérieur à la séparation entre les plantes à fleurs (Angiospermes) et les plantes à graines nues (Gymnospermes), et un second vers 200-230 millions d'années (Ma), commun à tous les Angiospermes (Jiao *et al.* 2011) (Figure 1). L'évaluation du nombre de duplications ancestrales de gènes attribuées à des WGD le long d'arbres phylogénétiques a permis d'identifier approximativement 50 événements de polyploïdisation au sein des lignées végétales (Zhang *et al*. 2019).

La fréquence d'occurrence des évènements de polyploïdisation chez les angiospermes a été largement étudiée mais selon la méthode utilisée et les familles d'angiospermes sélectionnées, les résultats diffèrent sensiblement : 35% (Wood *et al*. 2009, phylogénie et cytogénétique), 47% (Grant 1980, nombre de chromosome dans les cellules haploïdes de base des familles étudiées n=14), 70% à 80% (Goldbatt 1980, n=9 ou 10), 70% (Masterson 1994, taille des cellules de garde des stomates sur espèces fossiles et actuelles de trois familles d'angiospermes).

La récurrence des évènements de polyploïdisation chez les végétaux interroge quant à l'avantage adaptatif de ce phénomène et son impact sur le fonctionnement et l'évolution des génomes. L'enjeu des questions de recherche sur la polyploïdie, nécessite de comprendre dans un premier temps les mécanismes de formation des espèces polyploïdes et les bouleversements génomiques qui font suite.



I.2 Formation des espèces polyploïdes

Plusieurs voies de duplication intégrale d'un génome (WGD) ont été identifiées pour conduire à un individu présentant un génome polyploïde (Tayalé et Parisod 2013, Soltis *et al.* 2014, Ramsey et Schemske 1998). Les espèces polyploïdes sont classées en deux principales catégories selon leur mode de formation (Figure 2).

Figure 2. Les voies naturelles de formation d'un individu polyploïde.

La voie de doublement somatique du génome est représentée en pointillés car elle est considérée comme rare. Les phases suivant la formation des individus néopolyploïdes subissent une phase de réorganisation intense du génome (schématisée par les changements de couleur des points représentant les sous-génomes) (Tayalé et Parisod 2013).

I.2.1 **Formation des autopolyploïdes**

D'un point de vue mécanistique, un individu autotétraploïde peut dériver :

- d'un doublement chromosomique somatique au niveau du zygote à la suite d'une erreur post-réplication de l'ADN au cours des premières mitoses
- de la fusion de gamètes non réduits, eux-mêmes provenant d'une erreur post-réplication de l'ADN au cours de la méiose chez un ou plusieurs individus de la population

Les espèces autopolyploïdes » (pomme de terre, canne à sucre, luzerne) présentent alors plusieurs jeux de chromosomes provenant d'une seule et même espèce et les chromosomes sont alors considérés comme pratiquement « homologues » car très similaires. La construction génomique symbolique de ces espèces peut s'écrire : AAA'A'.

I.2.2 **Formation des allopolyploïdes**

La formation d'individus dits allopolyploïdes s'explique par une hybridation interspécifique chez des individus de deux espèces distinctes mais proches géographiquement, permise par une synchronisation des stades d'organogénèse florale et de production des gamètes. Dans ce cas, l'individu allopolyploïde provient théoriquement de :

1. La fusion de deux gamètes non réduits de deux espèces différentes produits simultanément par deux espèces (rare)
2. La fusion de deux gamètes haploïdes de deux espèces différentes puis doublement chromosomique au niveau du zygote hybride
3. Pont triploïde avec fusion d'un gamète non réduit avec un gamète haploïde puis doublement du génome

Les espèces **allopolyploïdes** dont le génome comprend des jeux de chromosomes provenant d'au moins deux espèces différentes ; leurs chromosomes et les sous-génomes sont alors considérés comme homéologues (coton, blé, colza, caféier, tabac, arachide, choux, …) ; construction génomique : AA'BB' pour une espèce tétraploïde.

La production de gamètes non réduits est présentée comme un processus courant chez les végétaux et sans doute à l'origine de la formation de nombreuses espèces polyploïdes (Bretagnolle et Thompson 1995). La possibilité de fusion simultanée de deux gamètes non réduits étant considérée comme peu probable, il a été admis que la formation d'individus autotétraploïdes passait certainement plus souvent par une voie indirecte appelée « **pont triploïde** » (*Unilateral pathway* Figure 2, Comai 2005). Il s'agit d'un individu obtenu à partir de la fusion d'un gamète non réduit et d'un gamète haploïde, considéré comme étant un intermédiaire entre des individus diploïdes et tétraploïdes. De même, certains individus appelés « **pont homoploïde** » (*homoploïd bridge*, Figure 2) peuvent conduire à la

formation d'individus allopolyploïdes (Behling *et al*. 2019). Ils sont issus d'une hybridation de gamètes haploïdes sans doublement des chromosomes chez le zygote directement après l'hybridation du génome. Une des particularités de l'allopolyploïdisation est la possibilité d'une hybridation entre espèces ne possédant pas le même nombre de chromosomes dans leurs noyaux. C'est le cas de *Brassica napus* qui est issu d'une hybridation entre *Brassica oleracea* (2n=18) et *Brassica rapa* (2n=20).

Bien qu'il soit commode de distinguer deux types majoritaires de forme de polyploïdes, certains individus polyploïdes présentent un intermédiaire entre les deux types pré-cités, issus d'un doublement de génomes partiellement différentiés (Tayélé et Parisod 2013). Deux autres catégories de polyploïdes ont été observées mais sont plus rares : les autoallopolyploïdes (construction génomique AA'BB'BB', exemple *Héliantus tuberosus*) et les allopolyploïdes segmentaires (A1A1A2A2) qui correspondent à une mosaïque de sous-génomes similaires mais pas identiques (Dar et Rehman 2017).

La formation d'un nouvel individu polyploïde entraîne de nombreux changements génomiques, au niveau de la structure du génome avec des réarrangements chromosomiques et au niveau de son fonctionnement avec des remaniements de la régulation de l'expression des gènes. Ces changements se produisent de façon plus ou moins progressive ou instantanée, selon qu'ils sont étudiés à des échelles de temps post-polyploïdisation plus ou moins longs.

I.3 Conséquences génomiques de la polyploïdisation

I.3.1 La polyploïdisation, phénomène cyclique ?

Plusieurs analyses paléogénomiques résumées dans les revues de Zhang *et al*. 2019, Wendel 2015, Baduel *et al*. 2018 ont démontré que le nombre d'événements de WGD dans les lignées végétales et (1) le nombre de chromosomes (2) et la taille des génomes sont deux paramètres qui ne semblent pas corrélés (Baduel *et al*. 2018). En effet, on pourrait s'attendre à une augmentation du nombre de chromosomes et de la taille des génomes dans les lignées évolutives présentant des évènements de polyploïdisation. Bien que la taille des génomes s'explique également, par les évènements de recombinaison (entrainant des délétions) et les évènements de transposition des éléments transposables (entrainant une accumulaiton d'ADN), ce n'est pas ce qui est observé. Par exemple, les plantes du genre *Brassica* qui ont subi trois événements de WGD (α, β, γ, Franzke *et al*., 2011; Jiao *et al*., 2011) au cours de leur histoire évolutive, devraient posséder théoriquement 40 à 56 chromosomes si les estimations théoriques du nombre de chromosomes des espèces ancestrales sont justes (cinq ou sept chromosomes, Jiao *et al*. 2011). Pour l'espèce de coton *Gossypium hirsutum*, de fait des cinq ou sept événements de WGD ayant conduit à cette espèce (Renny-Byfield *et al*. 2014), celle-ci devrait contenir actuellement dans son noyau haploïde 224 à 896 chromosomes (selon le principe de u(n+1) = u*2 avec u (0) = 7 ; u=nombre de chromosomes initial) et présenter une taille de 43,2 Gb

(actuellement 2,4 Gb). Ceci implique que l'état polyploïde n'est pas maintenu en tant que tel mais que le nombre de chromosomes et la taille du génome diminuent au cours de l'évolution.

Le concept décrivant ce phénomène a été appelé diploïdisation post-polyploïdisation (PPD, Post-Polyploïdization Diploïdisation) évoqué plus haut dans cette synthèse. Dans leur revue, Mandakova et Lysak 2018 parlent de trois catégories d'espèces polyploïdes, classés selon leur âge de formation : néo-polyploïdes (*Arabidopsis sueca, Tragopogon*, blé tendre), méso-polyploïdes (*Brassica rapa*) et paléo-polyploïdes (*Arabidopsis thaliana*, maïs, soja). Ces différents génomes polyploïdes sont caractérisés par un avancement différentiel du phénomène de PPD (Diploïdisation Post-Polyplïdisation) dont l'appariement des chromosomes en méiose et le nombre de chromosomes sont les témoins. Plus l'héritabilité des chromosomes en méiose se rapproche de celle d'un individu diploïde (héritabilité disomique), plus le nombre de chromosomes de l'espèce polyploïde se rapproche de celui de l'espèce ancestrale diploïde, et plus l'espèce polyploïde est considérée comme ancienne.

L'interprétation des analyses paléogénomiques et cytogénétiques du nombre et de la nature chimérique des paléochromosomes a conduit certains chercheurs à penser les processus de polyploïdisation/diploïdisation comme des phénomènes cycliques : un génome polyploïde ne se perpétuerait pas en tant que tel mais serait voué à retourner vers un état diploïde (Figure 3 ci-dessous).

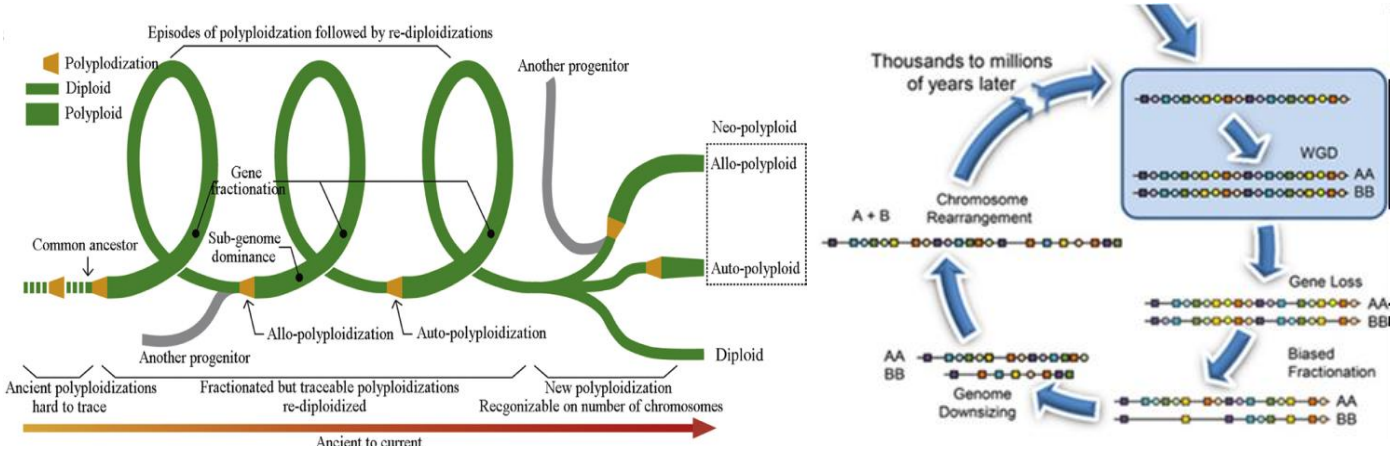


Figure 3. Schémas traduisant la nature cyclique et récursive du processus de polyploïdisation et Diploïdisation Post-Polyploïdisation.

A gauche, schéma de Zhang et al. 2019 représentant l'histoire évolutive d'espèces allo ou autopolyploïdes.
A droite, schéma de Wendel et al. 2015 représentant les mécanismes associés au retour à la diploïdie (processus de Diploïdisation)

I.3.2 Remaniements génomiques dans les premières générations

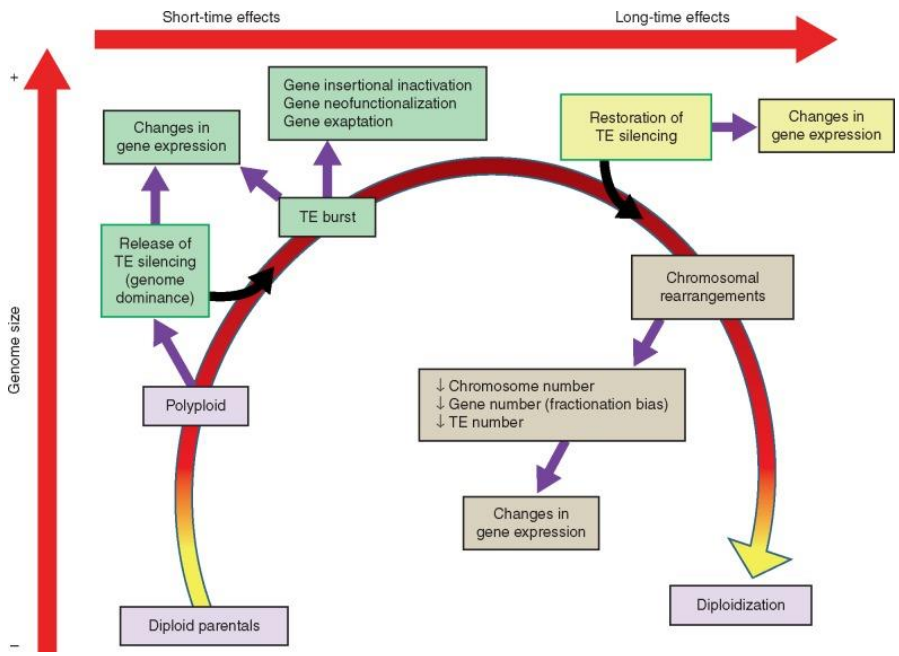
I.3.2.1 Stabilisation des chromosomes et maintien d'une méiose assurant la survie de l'espèce

Selon les modes de formation des polyploïdes, les chances de survie sont en parties liées à la stabilisation des chromosomes pour les mitoses et à la réussite du processus de méiose (Pelé *et al.* 2018). En effet, l'appariement des chromosomes homologues, grâce à l'établissement de crossing-over, est nécessaire pour assurer une ségrégation correcte des chromosomes durant la division

équationnelle de méiose. Un défaut dans cet appariement (appariement entre chromosomes homéologues par exemple) peut entraîner un défaut du nombre de chromosomes dans les gamètes (aneuploidie) et compromettre les chances de survie de l'individu néopolyploïde. Chez les espèces autopolyploïdes, les chromosomes sont très similaires en termes de séquence ce qui favorise la formation de multivalents : association de plus de deux chromosomes au cours de la méiose grâce aux chiasmas qui correspondent à la réalisation des crossing-overs, ce qui peut conduire à des gamètes aneuploïdes (mauvaise séparation des chromatides). Chez les espèces allopolyploïdes, les séquences entre chromosomes homéologues sont souvent suffisamment divergentes pour distinguer les chromosomes homologues et homéologues. On observe alors un appariement « bivalent » (par paires) lors des méioses dès les premières générations mais cela dépend du degré de divergence de séquence des génomes des espèces diploïdes progénitrices (Pelé *et al.* 2018).

Ainsi, l'émergence et la survie d'un grand nombre de polyploïdes pourraient être attribuées au fait que ces espèces possèdent une certaine plasticité des mécanismes de méioses. Plusieurs auteurs ont notamment montré une fréquence de crossing-overs plus importante chez les individus polyploïdes (*Arabidopsis*, *Brassica*, *Gossypium* et *Zea)*, comparativement à leurs progéniteurs (Pelé *et al.* 2017). La polyploïdisation peut également favoriser des évènements de recombinaisons illégitimes entre chromosomes homéologues (NonAllelic Homoeologous REcombination (NAHR)), qui peuvent être nombreux dans les premières générations, et entraîner ainsi l'apparition d'échanges de fragments d'ADN appelés « homoeologous exchanges ». Ces évènements peuvent avoir une incidence sur les variations du nombre de copies de gènes par exemple (Loyd *et al*. 2018), ou sur les variations de traits quantitatifs (Stein *et al.* 2017). Rousseau-Gueutin *et al*. 2016 ont notamment montré que 10% des gènes étaient impactés par des variations du nombre de copies dues à des recombinaisons illégitimes après seulement trois générations chez l'espèce synthétique allotetraploïde *Brassica napus*.

I.3.2.2 **Remobilisation des éléments transposables**



Un autre phénomène pouvant entraîner de nombreux réarrangements de séquences dans le génome polyploïde réside dans la réactivation de familles d'éléments transposables (TE, Transposable Elements). Cette réactivation a été observée chez quelques espèces de polyploides resynthétisés ou

Figure 4. Schéma des liens entre polyploïdisation et Eléments Transposables (Vicient et al. 2017).

de neo-polyploides naturels mais concerne seulement quelques familles de TEs : *Nicotiana* (Petit *et al.* 2010), *Brassica* (Alix *et al.* 2008), *Gossypium* (Grover *et al.* 2008). A contrario, il est possible que l'événement de polyploïdisation ne réactive aucune famille d'éléments transposables comme cela a été observé chez des colzas resynthétisés (Sarilar *et al.* 2013). L'impact de la transposition sur l'évolution des génomes végétaux et notamment de la partie codante du génome, par l'acquisition de nouvelles séquences régulatrices est notamment résumée dans la revue de Vicient *et al.* 2017 (Figure 5). Cet aspect ne sera pas plus développé dans cette introduction mais il faut savoir que les changements d'environnement en TE au niveau des séquences régulatrices des gènes à la suite d'un événement de polyploïdisation peuvent remanier l'expression de certains gènes et ainsi être à l'origine de l'apparition de phénotypes nouveaux.

I.3.3 **Remaniements génomiques progressifs lors de l'évolution des espèces polyploïde**

I.3.3.1 **Perte de séquence et fractionnement du génome**

I.3.3.1.1 **Mécanismes associés à la perte de séquence**

La comparaison de génomes paléopolyploïdes et néopolyploïdes a conduit à observer une réduction de la taille des génomes polyploïdes au cours de leur évolution *via* la diminution du nombre de chromosomes (« descending dysploïdy ») et la perte de séquence (« genome down sizing ») (Mandakova et Lysak 2018). L'ensemble de ces phénomènes fait parti du processus de Diploïdisation Post-Polyplïdisation. Le premier phénomène ne sera pas évoqué ici et seuls les mécanismes entraînant la perte de séquences seront décrits. Ce phénomène, appelé « fractionnement du génome » (« genome fractionation » en anglais), correspond à l'élimination des copies homéologues redondantes issues d'un événementde WGD (Cheng *et al.* 2018). Cette perte se produit essentiellement au niveau des séquences répétées mais peut également toucher les gènes selon différents mécanismes.

A. Perte de séquences répétées

Une perte de fragments de séquences peut se produire par réarrangements chromosomiques et recombinaisons illégitime notamment au niveau des séquences répétées de type éléments transposables (Mandakova and Lysak 2018, Vicient and Casacuberta *et al*. 2017). La revue de Vicient et Casacuberta (2017) résume très bien les mécanismes pouvant expliquer une perte importante de séquences avec en particulier la recombinaison liée à la réparation de l'ADN et l'abondance des TEs. Des exemples de perte de séquences par le biais de recombinaisons et de réarrangements chromosomiques ont été décrits chez le tabac (Lim *et al*. 2007) et le maïs (Bruggmann *et al*. 2006).

B. **Perte de séquences codantes**

La perte progressive de séquences codantes fonctionnelles, qui étaient maintenues par pression de sélection chez les parents diploïdes, mais qui sont alors devenues redondantes chez l'espèce polyploïde peut être expliquée par trois principaux mécanismes :

a) Via une délétion non contre sélectionnée due à un événement de cassure de l'ADN mal réparé ou à la suite de l'excision d'un TE de type rétrotransposon (Vicient et Casacuberta (2017).

b) Par dérive génétique lié à des mutations aléatoires : la séquence du gène ou sa séquence régulatrice en amont vont accumuler des mutations aléatoires à la suite du relâchement de la pression de sélection dû à la redondance génétique. L'une des copies homéologue sera plus touchée aléatoirement et va accumuler des mutations non contre-sélectionnées qui vont peu à peu fragmenter la séquence (pseudogénisation) (Meirmans *et al.* 2018).

c) Par dérive génétique liée à des divergences d'expression des copies de gènes (Bottani *et al*. 2018)

- Si la régulation de l'une des copies chez l'espèce diploïde s'avère plus répressive comparativement à la séquence de l'autre parent progéniteur. En effet, si cette régulation est conservée alors l'une des copies sera moins transcrite chez le polyploïde et elle sera potentiellement moins soumise à pression de sélection et pseudogénéisée
- Par évolution fonctionnelle différentielle des copies homéologues : l'une des copies peut acquérir une régulation répressive à la suite de la polyploïdisation. Elle sera alors moins indispensable au fitness des individus et sera également progressivement pseudogénéisée

En ce sens, un modèle appelé « use it or lose it » (utilisé sinon perdu) faisant appel à des processus épigénétiques (Figure 5 ci-dessous) a été proposé (Bottani *et al*. 2018). Selon ce modèle certains gènes dupliqués vont présenter une régulation qui va peu à peu devenir défavorable à l'expression et accélérer l'élimination de ces séquences par dérive génétique.

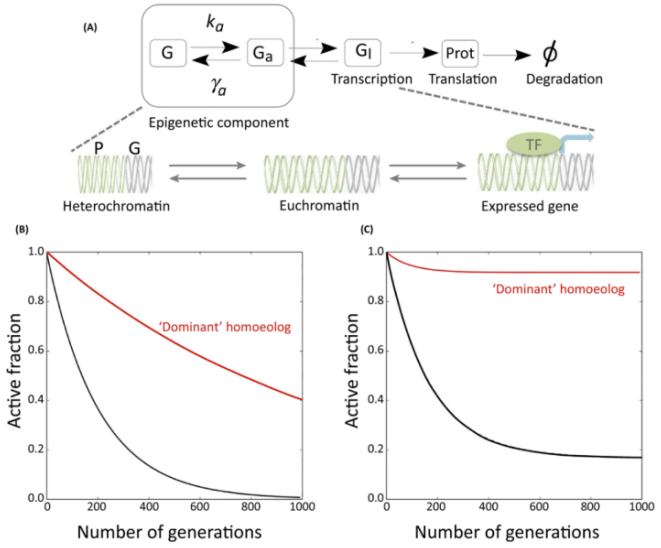


Figure 5. Processus d'inactivation épigénétique de séquences homéologues selon le modèle de Bottani et al. 2018.

*Au départ les deux séquences G et G' peuvent être toutes les deux initialement accessible (G_a) à la machinerie de transcription puis de façon stochastique devenir inaccessibles (G) ou induite (G_I) **(a)**. Les constantes K_a et γ_a représentent l'accessibilité ou inaccessibilité chromatinienne. Or, si k_a=0 l'accessibilité de la copie homéologue la plus faiblement exprimée peut diminuer à la suite de la réduction de l'induction/activation. Au cours des générations, à l'échelle individuelle ou populationnelle, il peut se produire une « extinction » de cet homéologue récessif dans une majorité de cellules/individus **(b)**. Si ka>0, les conditions épigénétiques sont réversibles et l'expression de l'homéologue récessif peut se maintenir **(c)**.*

I.3.3.1.2 Tendances observées pour le fractionnement des génomes

L'analyse de plusieurs génomes polyploïdes anciens (paléopolyploïdes) a permis de discerner deux tendances majeures quant au fractionnement du génome. Il peut se faire de façon équilibrée entre les deux sous-génomes ou biaisée envers l'un ou l'autre « biaised fractionation en anglais » (Liang et Schnable 2018, Bird *et al.* 2018, Figure 6 ci-après).

A. Fractionnement biaisé

Dans le cas d'un fractionnement biaisé, la perte de séquence peut être reliée au différentiel de transcription global entre les deux sous-sets de gènes homéologues. Si l'un des deux sous génomes présente des niveaux d'expression plus faible, la pression de sélection sera moins importante sur ces gènes moins exprimés entraînant une perte progressive par pseudogénisation biaisée des séquences codantes. C'est ce que Schnable *et al*. 2011a avaient démontré pour le paléopolyploïde *Zea mays* avec un fractionnement plus important du sous-génome 2 qui présentait un niveau d'expression des gènes globalement plus faible. De même, l'analyse de l'évolution des sous-génomes de *Brassica rapa* par Cheng *et al*. 2012 avait mis en évidence un différentiel de pression de sélection sur les gènes homéologues présentant un biais d'expression, avec une proportion de gènes perdus moins importante pour le sous-génome dont le niveau d'expression dominait.

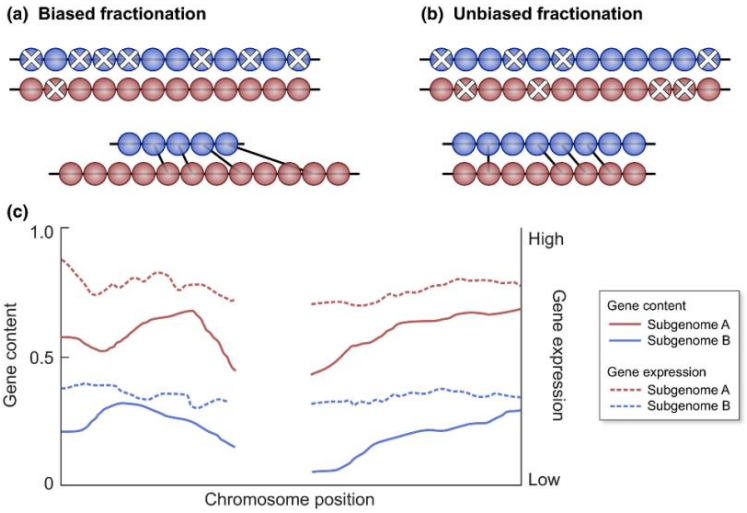


Figure 6. Schéma des patrons de fractionnement des génomes polyploïdes.

a) *et b) représentent les pertes de gènes non-équilibrée ou équilibrée entre les deux sous génomes respectivement. c) Schéma de la perte de gènes reliée à l'expression des gènes à l'échelle des chromosomes.*
Bird et al. 2018

B. Fractionnement « équilibré »

Dans le cas d'un fractionnement équilibré entre les deux sous génomes, les différentiels de transcription entre gènes homéologues dépendent du tissu, de l'organe ou du stade de développement. Ces différences sont liées à des divergences stochastiques entre les niveaux d'expressions des gènes chez les génomes diploïdes progéniteurs. Ainsi, en fonction des niveaux de transcription des homéologues de l'un des sous-génomes dans une condition particulière, la pression de sélection va maintenir la copie exprimée de façon la plus optimale (niveau, dynamique…). Cela correspondra à un fractionnement « mosaïque ». En comparant le fractionnement des génomes de deux espèces paleopolyploïdes : le soja et maïs (polyploïdisation : entre 5 - 13 MYA et 11,5 MYA respectivement), Zhao et collaborateurs (2017) ont démontré une évolution contrastée en termes de perte de séquences pour les deux espèces. L'un des sous génomes du maïs présentait une expression globale plus faible, une pression de sélection moins marquée et un fractionnement plus important. *A contrario*, chez le soja ce n'est pas un fractionnement biaisé qui a été observé une rétention des gènes dupliqués beaucoup plus importante et des pertes de séquences stochastiques entre les deux sous-génomes. Les auteurs expliquent ce contraste par une divergence marquée des espèces progénitrices du maïs qui n'est pas retrouvée chez celles ayant donné le soja. Le maïs serait issu d'une allopolyploïdisation de génomes diploïdes déjà fortement différenciés ante-polyploïdisation.

Selon les travaux de Garsemeur *et al*. 2012, dont l'étude portait sur l'analyse des blocs de synténie (groupes de gènes dont le voisinage et l'organisation sont conservés entre plusieurs génomes, avec des relations de colocalisation et d'identité de séquences) chez le peuplier, le soja, la luzerne, Arabidopsis, le sorgho, les brassica et le maïs, les espèces polyploides pourraient être classées en trois catégories :

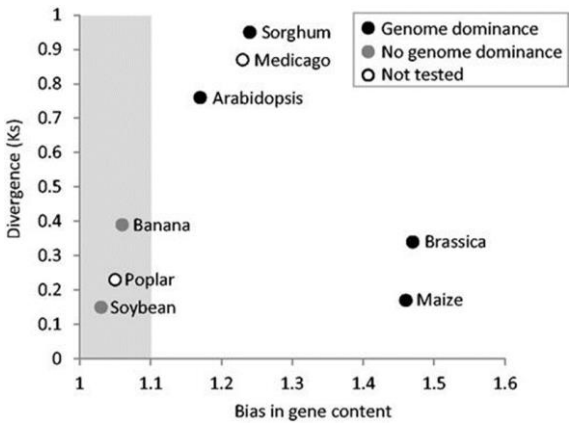


Figure 7. Classification d'espèces polyploïdes (pertes de gènes et divergences génétiques).

L'axe des abscisses représente la médiane des biais de contenu en gènes entre sous-génomes et l'axe des ordonnées représente les taux de mutations de substitution entre gènes homéologues résultat de la plus récente WGD dans l'espèce considérée. L'aire grise sur le diagramme ou figure la banane comprend une WGD n'incluant pas de fractionnement biaisé ni de dominance de sous génomes entre les génomes dupliqués. Garsemeur et al. 2013

- Les espèces présentant une dominance d'expression de l'un des sous-génomes accompagné d'un fractionnement biaisé des gènes : maïs, brassica (allopolyploides)
- Les espèces présentant une expression équivalente des sous-génomes et une perte de gène loci-dépendante : poire, soja, banane (autopolyploides)
- Des espèces intermédiaires avec une dominance globale d'expression mais un fractionnement biaisé des séquences géniques modéré : sorgho, Arabidopsis

La Figure 7 issue de leur publication résume la classification proposée selon les paramètres de dominance d'expression, de fractionnement biaisé de l'un des sous génomes et de divergence de séquence entre les espèces étudiées. Les auteurs avaient conclu que le type de formation des polyploïdes (auto vs allo polyploïdization) pouvait être le paramètre expliquant cette classification. A l'inverse de la perte de séquence observée chez tous les polyploïdes, la rétention de séquences dupliquées est aussi un phénomène majeur dans l'évolution des génomes polyploïdes.

I.3.3.2 Maintien des copies de gènes dupliqués : évolution fonctionnelle du génome

Plusieurs mécanismes ont été identifiés pour expliquer le phénomène de la rétention préférentielle de certains gènes dupliqués en plusieurs copies :

- **Processus de sous-fonctionnalisation** de l'un des copies dupliquées : mutation et évolution de la séquence le plus souvent au niveau de des éléments de régulation en *cis* avec conservation de la fonction ancestrale pour l'un des copies et évolution pour l'autre copie (Roulin *et al.* 2013, Freeling *et al.* 2015, Cheng *et al.* 2018)

- **Processus de néo-fonctionnalisation** : mutations ayant lieu au niveau de la séquence codante de l'une des copies dupliquées entraînant l'apparition d'une protéine avec une nouvelle fonction (Roulin *et al.* 2013, Hughes *et al.* 2014, Cheng *et al.* 2018)
- Gènes présentant une **expression dosage dépendante** (l'expression d'un nombre de copies de même fonction permettra de produire la quantité de protéines adéquate). Chez les polyploïdes, cette augmentation de la quantité de protéines ayant une même fonction explique souvent l'augmentation de la taille d'un organe ou l'efficacité d'un métabolisme, impliquant que le nombre de copies soit conservé (Raju 2020) ; cela correspond à l'hypothèse de l'équilibre de dose des gènes (« gene dosage balanced hypothésis) formulée par Birchler et Veitia 2012.

L'étude de ces mécanismes de rétention de gènes nécessite de différencier les gènes dupliqués par WGD de ceux dupliqués par d'autres mécanismes de duplications de gènes internes au génome : duplications en tandem, duplications proximales intrachromosomiques, duplications interchromosomiques. En effet, selon le type de duplication et les biais d'expression déjà présents chez les espèces diploïdes, les séquences dupliquées n'auront pas le même destin évolutif (Qiao *et al.* 2019, Conant *et al.* 2014).

1. Lors d'une duplication WGD, deux gènes redondants peuvent par exemple donner un phénotype plus favorable à travers l'additivité de leur expression et être tous les deux conservés. S'il existe une expression biaisée ou une efficacité fonctionnelle différentielle, l'un des deux gènes peut alors être éliminé par sélection « purificatrice » (Qiao *et al.* 2019).
2. Pour les gènes dupliqués par des mécanismes de type « single gene duplication », la conservation des deux copies dépendra 1) de la vitesse de divergence des deux copies ou 2) d'un potentiel effet dose pour la production de protéines qui apparaitrait à la suite de la duplication et qui serait sélectionné (Conant *et al.* 2014, Freeling 2009).

Selon Freeling *et al.* 2009, si deux gènes sont dupliqués en tandem dans l'un des ancêtres, il est probable que le couple de paralogues dupliqués en tandem soit conservé, et que le gène homéologue provenant de l'événementde WGD soit progressivement éliminé. Ainsi, potentiellement, le phénomène de fractionnement et de diploïdisation du génome polyploïde peut entraîner un déséquilibrage du dosage de gènes dupliqués par WGD favorisant la rétention de gènes dupliqués en tandem.

Dans leur revue, Conant *et al.* 2014 énumèrent les études ayant identifié certaines des fonctions des gènes dupliqués préférentiellement retenus : des gènes codant des protéines formant des complexes protéiques comme les ribosomes, les facteurs de transcription ou encore les protéines du protéasome. *A contrario,* les gènes moins souvent retenus sous forme de dupliqués (singletons) seraient des gènes dont la fonction ne nécessite pas de partenaire moléculaire (protéines du métabolisme de l'ADN, nucleases, protéines qui se lient à l'ARN). De façon générale, un grand nombre de gènes dupliqués

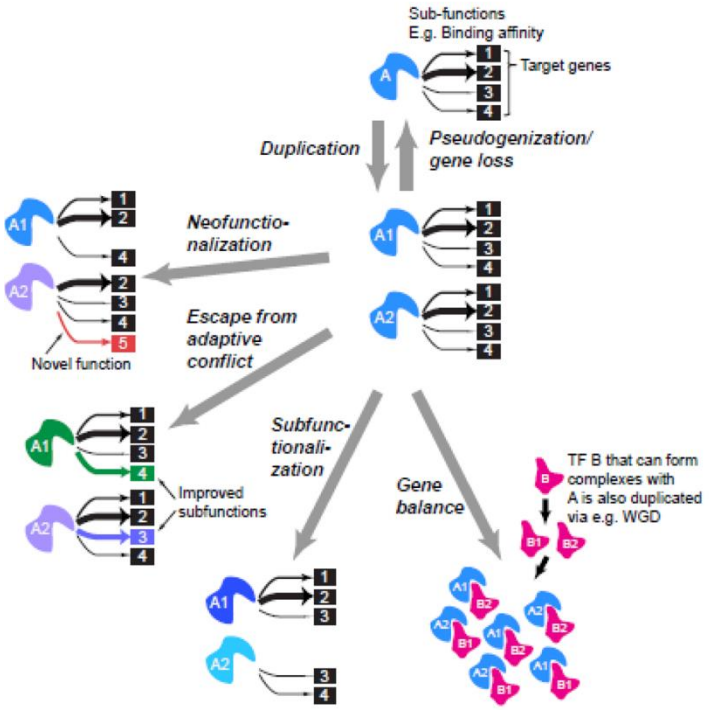


Figure 8. Destins évolutifs des facteurs de transcription dupliqués.

Lehti-Shiu et al. 2016

retenus à travers les différents événements de WGD et DPP correspondent à des gènes contrôlant des réseaux de régulation. L'étude de Lehti-Shiu *et al*. 2016 montre en particulier la diversité et l'expansion massive des gènes codant des facteurs de transcription chez les plantes et leurs patrons évolutifs (Figure 8). Ceci pourrait expliquer cette incroyable richesse de formes et ce potentiel adaptatif pour ces êtres vivants fixés très fortement impactés par les conditions climatiques.

Pour résumer, dans leur étude Jiang *et al*. 2013 ont classé les gènes dupliqués retenus en trois catégories :

- Les gènes codant des protéines fonctionnant en complexes macromoléculaires (effet dose est important)
- Les gènes codant les facteurs de transcription, plus enclins à une possible sous-fonctionnalisation
- Les gènes de réponse aux stimuli environnementaux

Ainsi, les éléments influençant la rétention de séquences dupliquées au sein d'un génome, sont : la nature de la duplication, la divergence de séquence et de régulation entre les espèces progénitrices, l'effet dose, la fonction des gènes et les processus d'évolution des fonctions (neo/sous-fonctionnalisation des gènes).

I.4 Evolution fonctionnelle d'un génome allopolyploïde : régulation et expression des gènes

Que se passe-t-il lorsque deux génomes ayant divergé par spéciation et donc possédant un fonctionnement légèrement différent (régulation du développement, adaptation aux contraintes environnementales) se retrouvent dans un seul et même noyau ? Potentiellement, les gènes des deux génomes possèdent des éléments de régulations, des environnements en TE, des dynamiques de marquage épigénétique différents entraînant des niveaux, des amplitudes et des synchronisations d'expression des gènes différents et décalés. Ainsi, chez un allopolyploïde, l'enjeu pour la survie des

premières générations se situe dans la sélection d'individus dont l'expression des gènes permettra le développement d'un phénotype adapté et présentant des avantages sélectifs.

I.4.1 Inférence des gènes homéologues

L'étude des génomes polyploïdes et de n'évolution des gènes nécessite de connaître les relations évolutives entre les différentes copies de gènes. En effet, comme nous l'avons vu, les gènes dupliqués peuvent avoir plusieurs origines. Glover *et al*. 2016 propose une terminologie précise pour les différencier (Figure 9) :

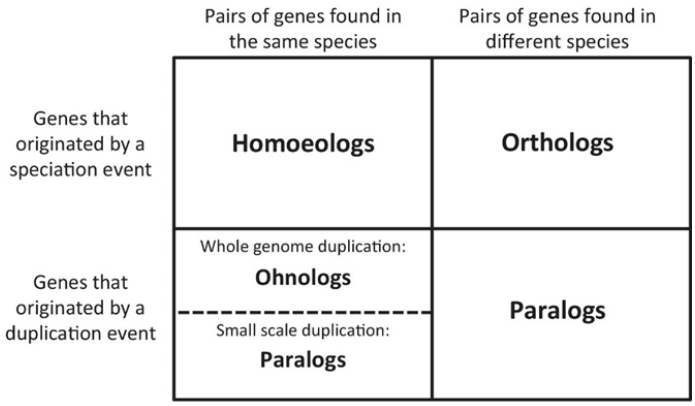


Figure 9. Les différentes catégories de gènes au sein d'un génome polyploïde.

Glover et al. 2016

1. « **Ohnologues** » : gènes dupliqués via autopolyploïdisation,
2. « **Homéologues** » : gènes dupliqués via allopolyploïdisation, définis comme ayant pour origine un gène ancêtre commun qui aurait évolué et divergé en séquence au cours de la spéciation de deux espèces et qui se retrouvent ensuite réunis dans le génome d'une espèce allopolyploïde
3. « P**aralogues** » pour les autres types de gènes dupliqués, notamment ceux issus de duplications de petite taille, intra ou inter-chromosomique, appelés en anglais « small scale gene duplication » (exemple duplications en tandem).

Les relations d'homéologie et de paralogie sont particulièrement complexes à reconstruire au sein d'un génome polyploïde. Ceci est d'autant plus vrai que la synténie (conservation de l'ordre de gènes le long des chromosomes homéologues) a été rompue à la suite de délétions, inversions, translocations ou autres réarrangements chromosomiques. Une méthode est majoritairement utilisée à l'ère du séquençage des génomes pour l'inférence des gènes homéologues et paralogues chez les espèces polyploïdes (Glover *et al.* 2016). Elle est fondée sur la détection des relations évolutives en utilisant des séquences géniques et des arbres phylogénétiques des espèces. Deux gènes homéologues dans une espèce allopolyploïde sont définis comme des « orthologues » et leurs relations phylogénétiques sont analysées comme s'ils appartenaient à des espèces différentes. Ensuite, la méthode permet de déterminer les relations d'homéologie et de paralogie entre séquences en utilisant le processus de réconciliation des deux types d'arbres phylogénétiques (Figure 13). Si l'arbre des gènes suit l'arbre des espèces alors les deux gènes présentent des relations d'homéologie. Si ce n'est pas le cas, les deux types d'arbres sont réconciliés en utilisant le principe de parcimonie et en inférant une duplication intragénomique ayant donné un gène paralogue.

I.4.2 **Analyse de l'expression des gènes chez une espèce polyploïde**

L'article fondateur de Grover *et al*. 2012 explicite de façon très claire les deux méthodes distinctes d'analyse du transcriptome chez les espèces allopolyploïdes (figure 10 ci-après):

A. Analyse des biais d'expression entre gènes homéologues

Cela consiste en la comparaison de l'expression des gènes homéologues au sein du génome polyploïde actuel par calcul des contributions relatives de chaque homéologue à l'expression globale du locus. Cela permet d'étudier et comprendre les patrons d'expression des gènes homéologues sélectionnés au cours de l'évolution de l'espèce. On pourra construire un atlas de contributions différentielles entre sous-génomes par tissus et pour chaque groupe d'homéologues. On peut détailler ces analyses en rajoutant le paramètre du nombre de copies homéologues.

B. Analyse de la dominance d'expression des sous-génomes

Il s'agit ici de la comparaison de l'expression des gènes homéologues avec celle des gènes orthologues présents chez les espèces diploïdes parentales : analyse de l'additivité ou du caractère transgressif de l'expression de l'espèce polyploïde par rapport aux espèces progénitrices. Cela permet d'analyser la mise en place de nouveaux patrons de régulation post-polyploïdisation entre les espèces diploïdes et polyploïdes et l'évolution de l'expression des gènes homéologues dans le contexte nucléaire polyploïde. Cependant, ces analyses ne donnent qu'une vision biaisée des différences puisque les trois espèces ont continué à diverger et à évoluer chacune indépendamment. Ainsi, comparer des profils d'expression entre espèces diploïdes et polyploïdes recrées artificiellement (polyploïdes synthétiques) est plus informatif.

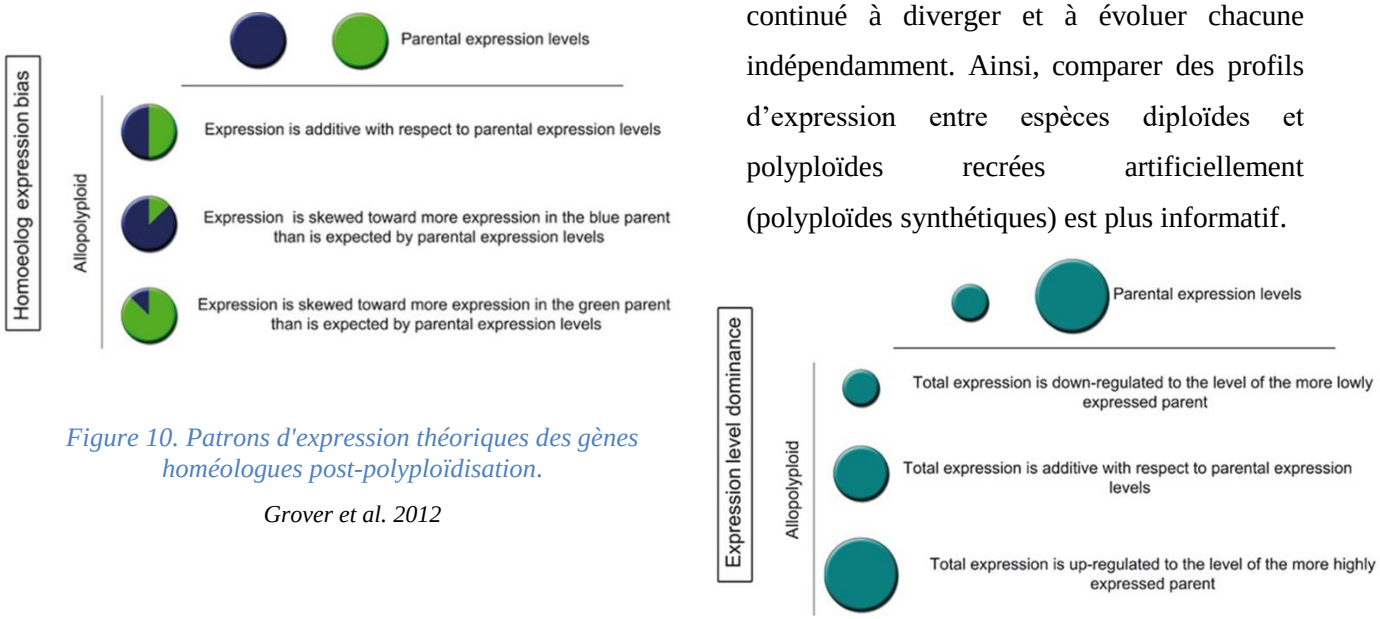


Figure 10. Patrons d'expression théoriques des gènes homéologues post-polyploïdisation.

Grover et al. 2012

I.4.3 **Patrons d'expression théoriques et observés**

D'un point de vue théorique, les analyses transcriptomiques doivent révéler des résultats sensiblement différents selon le type de polyploïdisation (autopolyploïdie ou allopolyploïdie) mais également selon le degré de divergence des génomes diploïdes progéniteurs (Comai 2005).

- Chez les espèces autopolyploïdes, présentant des séquences codantes et de régulation quasiment identiques, on devrait observer théoriquement une expression additive des deux sous-génomes
- Chez les espèces allopolyploïdes, le mélange de deux génomes peu ou très divergents peut conduire à une réorganisation plus ou moins marquée de la régulation transcriptionnelle et on devrait observer plus facilement des biais d'expression (Liang et Schnable 2018)

La Figure 11 ci-dessous, issue de la thèse de Marie-Christine Combes-Gavalda (2015) elle-même adaptée des analyses de Rapp *et al*. 2009, résume parfaitement les différents cas de figure observables lors d'une étude comparative des transcriptomes des espèces parentales diploïdes et celui de l'allopolyploïde :

1) **L'additivité de l'expression** : la somme de l'expression des deux homéologues (gènes ou génomes) au sein de l'individu polyploïde correspond à la somme de (en valeur absolue) des expressions mesuréees chez chacun des parents diploïdes
2) **La transgression d'expression** : le niveau d'expression d'un ou des deux sous-génomes homéologues chez l'individu polyploïde est supérieur ou inférieur à celui de l'un ou des deux parents, et ce quel que soit l'expression parentale de base (différentielle ou non entre les deux diploïdes).
3) **La dominance** : le niveau d'expression global de l'un des sous-génomes ou de gènes homéologues sont proches de celui de l'un des deux parents et supérieur à l'autre ce qui entraîne un déséquilibre d'expression au sein du génome polyploïde.
4) **Pas de changements** dans les niveaux de transcription, les niveaux d'expression des gènes des trois espèces sont égaux

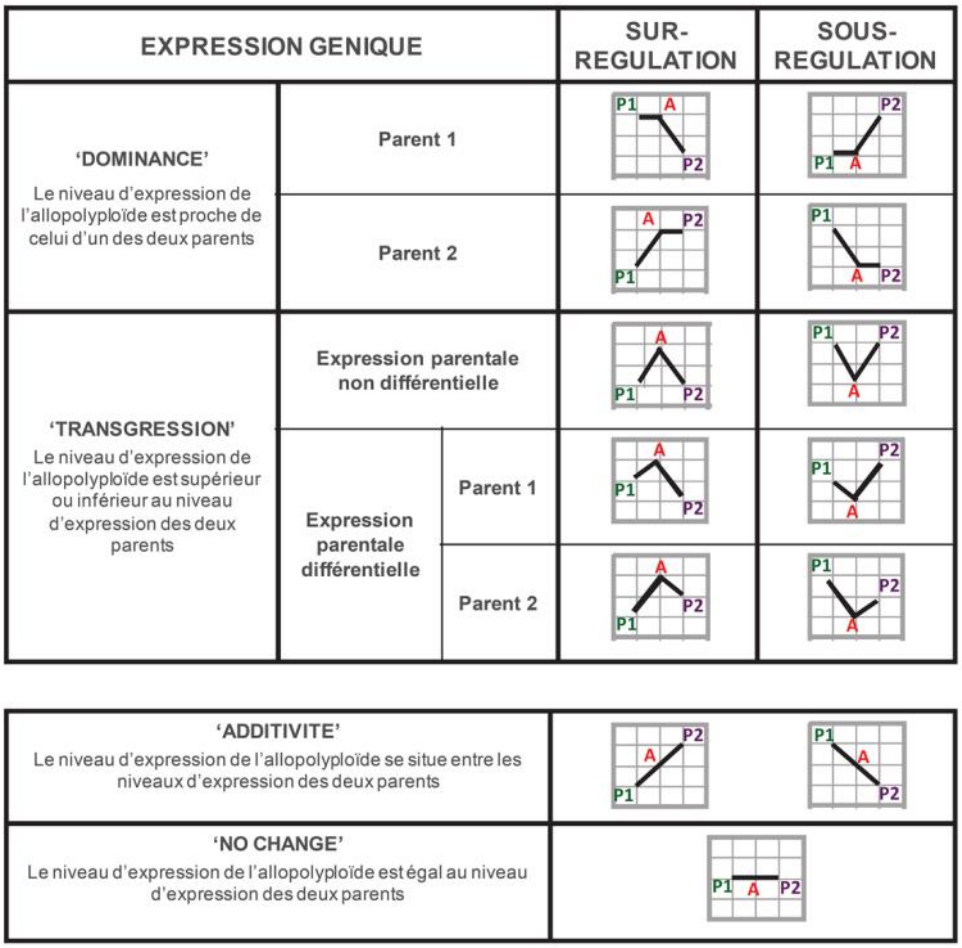


Figure 11. Profils théoriques d'expression de gènes post-polyploïdisation.

Marie-Christine Combes-Gavalda (2015)

I.4.4 Tendances observées pour l'expression des gènes homéologues chez différentes espèces allopolyploïdes

Un panorama des études réalisées chez différentes espèces polyploïdes permettra d'avoir un aperçu des tendances observées en termes d'expression des gènes en fonction de l'âge de formation de l'espèce polyploïde et du degré de divergence entre les espèces progénitrices.

Les études pionnières qui ont caractérisé les conséquences transcriptomiques d'une hybridation et d'un doublement de génome chez les espèces allopolyploïdes sont celles de Rapp *et al*. 2009 puis de Flagel et Wendel *et al*. 2010 sur le coton, Chelaifa *et al*. 2010 sur la spartine et Bardil *et al*. 2011 sur le café. Ces études ont toutes révélé une expression globale des gènes biaisée en faveur de l'un des sous génomes mais les données utilisées ne représentaient qu'une fraction des gènes. Plus récemment, plusieurs études portant sur la caractérisation des biais d'expression des gènes homéologues et l'évaluation de la dominance de l'un des sous-génomes ont été réalisées en utilisant d'individus synthétiques, des données RNA-seq sur plusieurs tissus et des systèmes biologiques d'étude comportant plusieurs espèces ayant conduit à l'événement de polyploïdisation.

I.4.4.1 L'exemple de Mimulus

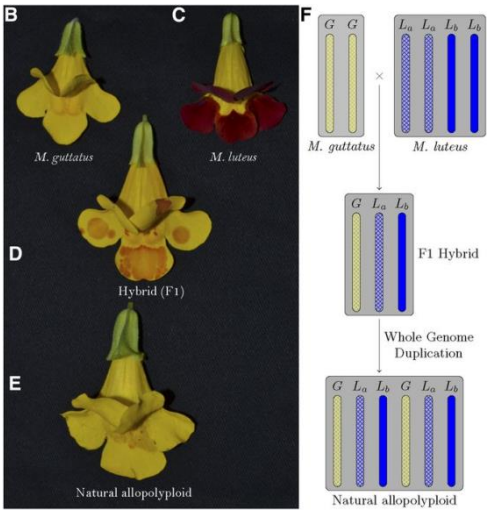


Figure 12. Système biologique des Lamiales du genre Mimulus, phénotype des fleurs et composition génomique.

B) à E) Espèces du genre Mimulus utilisées dans cette étude : M. guttatus (2x) (B) hybridé avec M. luteus (C) pour produire un triploïde stérile M. robertsii (3x) (D), lequel qui à une duplication de génome donne l'allopolyploïde naturel M. peregrinus (6x) (E). F) Graphique des génomes et sous-génomes de chaque espèce.

Le système expérimental et biologique proposé par Edger *et al*. 2017 chez Mimulus, avec les différentes formes sauvages disponibles, représentent un grand intérêt pour la compréhension des conséquences de la polyploïdisation sur l'expression des gènes (Figure 12). Les chercheurs ont travaillé sur le très récent allohexapolyploïde naturel *Mimulus peregrinus* (~140 ans) issus de l'hybridation entre une espèce diploïde *Mimulus gattatus* (2x) et une espèce tetraploïde *Mimulus luteus* (4x) portant les génomes L_a et L_b. Cette hybridation a conduit à la formation d'une espèce triploïde *M. x robertsii* stérile qui, par un doublement de son génome, a donné l'espèce allohexaploïde *Mimulus peregrinus* (6x). Dans leur étude, les auteurs ont comparé des données de RNA-seq provenant de trois tissus sur les espèces diploïdes, hybrides F1, allopolyploïdes naturelles et synthétiques. Ils ont mis en évidence l'établissement de la dominance du sous génome de *M. luteus* au sein des descendants polyploïdes, avec une prépondérance plus accentuée pour le sous-génome Lb, dès l'hybridation chez l'espèce hybride F1. Ils ont aussi observé une augmentation des biais d'expression entre gènes homéologues entre les deux individus allopolyploïdes (synthétique et naturel), donc au cours des générations.

I.4.4.2 L'exemple du Coton

Le coton cultivé comprend deux espèces allotetrapolyploïdes issues de l'hybridation de différentes espèces diploïdes. *Gossypium hirsutum* L. provient de l'hybridation des espèces *G. arboreum* et *G. raimondii* il y a environ 1-2 millions d'années. Yoo *et al*. 2013 ont étudié des données RNA-seq de feuilles chez les espèces de cotons diploïdes (progéniteurs modèles : *G. arboreum* (A2) et *G. raimondii* (D5), allotetraploïdes domestiqués (*Gossypium hirsutum*), individus synthétiques, et l'hybride F1 (A2 x D5). Ils ont observé une dominance globale préférentielle du sous-génome A sur le sous-génome D avec plus de gènes présentant une expression similaire au génome A diploïde (*G. arboreum*). Ils ont également mis en évidence que les biais d'expressions entre gènes homéologues chez le polyploïde sont expliqués par une conservation des niveaux d'expression des gènes des espèces progénitrices. Enfin, ils ont également observé un renforcement des biais d'expression depuis la formation de l'hybride jusqu'à la spéciation de l'allopolyploïde avec statistiquement plus d'expression transgressive des gènes chez les espèces allopolyploïdes naturelles (16%, 14,1% et 13, 8% des gènes pour l'hybride F1, l'allopolyploïde sauvage et celui domestiqué respectivement), rejoignant les résultats trouvés chez Mimulus. Ces résultats ont été

remaniés avec le séquençage du génome de cette espèce et une analyse « whole genome » des données RNA-seq (Zhang *et al.* 2015). Cette étude a révélé une évolution asymétrique des deux sous génomes A et D en termes de perte de séquence mais pas de dominance globale de l'un des sous-génomes en termes d'expression. L'étude des biais d'expression sur 35 tissus différents a révélé que 20 à 40% des paires d'homéologues présentent des biais d'expression alternativement en faveur du sous-génome A ou D en fonction du tissu considéré. Les gènes du sous-génome A davantage exprimés correspondaient pour beaucoup à des facteurs de transcription.

I.4.4.3 L'exemple du genre Brassica

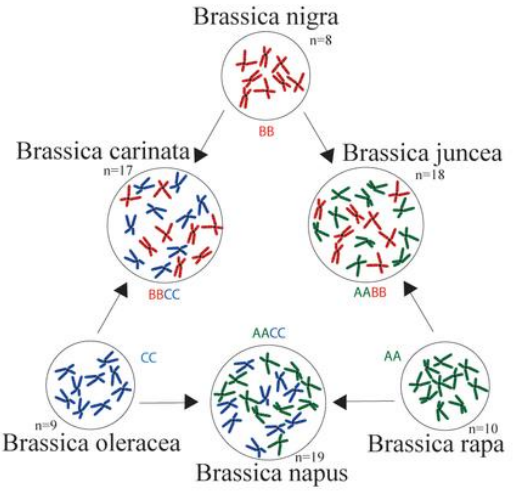


Figure 13. Triangle d'U, relations phylogéniques entre espèces du genre Brassica (Woo Jang-choon, 1935).

Les espèces du genre Brassica sont reliées entre elles par des évènements d'hybridation interspécifiques définies par le triangle d'U proposé par Nagaharu en 1935 (Figure 13). *Brassica napus* correspond au colza qui provient de l'hybridation récente entre *Brassica rapa* et *Brassica oleracea* qui a eu lieu il y a environ 7500 ans. Les patrons des biais d'expression chez cette espèce sont moins clairs que pour des allopolyploïdes plus anciens du fait de l'apparition très récente de cette espèce à l'échelle des temps évolutifs des espèces (7500 ans) et du fait des nombreux échanges de séquences entre sous-génomes qui peuvent avoir provoqué des remaniements transcriptomiques diluant les différences d'expression originelles entre les deux sous-génomes (Chahoub *et al.* 2014, Stein *et al.* 2017, Hurgobin *et al.* 2018). De nombreuses analyses d'individus de Brassica synthétiques ont été réalisées pour mieux comprendre le fonctionnement de ce génome (Zhang *et al.* 2010, Marmagne *et al.* 2010, Jiang *et al.* 2013, Tan et al 2016…). Par exemple, Wu *et al.* 2018 ont identifié 30% de gènes homéologues présentant une expression différentielle (Differential Gene Expression ou DEG) par rapport aux progéniteurs diploïdes avec 63% appartenant au sous génome C et 37% au sous génome D (d'après leurs travaux sur les 40371 gènes exprimés dans l'allopolyploïde synthétique). Les auteurs ont mis en évidence que les DEGs identifiées correspondaient à une répression de l'expression des gènes au sein de l'individu polyploïde comparée à l'expression des gènes parentaux. Ils ont également montré que 63,3% des paires de gènes homéologues présentaient des variations d'expression similaire à celles retrouvées chez les deux parents. Seuls 17% des gènes ont été identifiés comme présentant de nouveaux profils d'expression, différents des deux parents. Les auteurs ont ensuite caractérisé les biais d'expressions (additivité et transgressivité) globales chez le polyploïde synthétique comparé aux progéniteurs. Ils ont comptabilisé 46,5% de gènes ne présentant pas de changements de profils d'expression au sein du polyploïde, 4,9% d'expression additive, 9% d'expression transgressive et 36,9% présentant une dominance d'expression (ELD) avec significativement

plus de paires de gènes présentant un biais d'expression en faveur du génome A (24%) comparé au génome D (15%). Cette étude montre que chez une espèce polyploïde récente, il est difficile d'observer des dominances d'expression des sous-génomes et des biais d'expression entre gènes homéologues très importants ; les gènes homéologues semblent présentent pour une grande majorité une conservation de leur régulation à ce stade de l'évolution de l'espèce.

I.4.4.4 **Soja et Maïs, deux paléopolyploïdes à l'évolution contrastée**

Chez le maïs qui est une espèce paléopolyploïde, l'un des deux sous-génomes est soumis à une forte sélection négative (purifying selection) des allèles défavorables (Pophaly et Tellier 2015). Cette étude a également montré que 98% des paires de gènes paralogues retenus dans ce génome présentent une sous-fonctionnalisation de leur expression de façon tissu-spécifique. La plupart de ces gènes sont des facteurs de transcription alors que les gènes des paires de paralogues présentant une répression de l'expression sont des gènes de complexes macromoléculaires. Le soja est une espèce sont le génome a été modelé par deux évènements de polyploïdisation (59MYA et 13MYA) et présente 75% de ces gènes présents en copies multiples (Roulin *et al.* 2013). Parmi ces gènes, 50% d'entre eux présentent une sous-fonctionnalisation d'expression sur les sept tissus étudiés et correspondent en grande majorité à des facteurs de transcription. L'étude comparative des données RNA-seq entre les deux sous-génomes de ces deux espèces (Soybean 1 et Soybean 2 et Maïs 1 et Maïs 2) pour 24 et 28 conditions d'expression des gènes pour le maïs et le soja respectivement a mis en évidence une expression globale plus importante des gènes du sous-génome 1 pour le maïs (dominance de l'un des sous génomes) alors qu'aucune différence n'a été observée entre les deux sous-génomes du soja mais plutôt des biais d'expression stochastiques selon les tissus (Zhao *et al.* 2017). Ils ont corrélé les biais d'expression entre les deux sous-génomes du maïs avec une fréquence moins importante de TE proche des gènes dans le sous-génome maïs 1, bien que la proportion de TEs de ce sous-génome soit plus importante. Dans le génome du soja, aucune différence n'a été retrouvée concernant la proximité des TEs aux abords des gènes entre les deux sous-génomes. Le delta de divergences génomiques entre espèces progénitrices diploïdes des deux espèces étudiées a été évoqué pour expliquer les biais d'expression observés chez les polyploïdes actuels.

Cette liste de travaux n'est pas exhaustive et nombre de recherches sont effectuées chez d'autres espèces. Les travaux de Richard Buggs sur l'allotetraploïde *Tragopogon mirus* ont également révélé des tendances de biais d'expression des gènes liés à la sous-fonctionnalisation par tissus des gènes homéologues (Buggs *et al.* 2010, Buggs *et al.* 2012). Pour résumer, une des tendances observées chez beaucoup d'espèces allopolyploïdes est que la dominance de l'un des sous-génomes peut être expliquée par les caractéristiques structurales des régions régulatrices des gènes (présence et proportion de TE) déjà présentes chez les espèces progénitrices diploïdes. Des données d'expression sur plusieurs tissus permettent d'observer des sous-fonctionnalisations des gènes permettant la diversification de la redondance fonctionnelle et la rétention de gènes. L'effet dose d'expression lié au nombre de copies explique également la conservation des patrons d'expression de certains gènes.

Remarquons que les espèces polyploïdes très étudiées à l'heure actuelle sont des espèces domestiquées par l'homme. En effet, un grand nombre d'espèces domestiquées présente un état polyploïde ou palé-polyploïde de leurs génomes.

I.5 Polyploïdie et domestication des plantes

Les espèces dites domestiquées par l'homme correspondent à ces espèces non plus cueillies ou chassées et laissées à l'état naturel mais utilisées de façon récurrente dans les champs ou les jardins afin d'en récolter les fruits ou graines chaque année avec une idée de stockage des denrées récoltées. Ceci a par exemple impliqué de réserver certaines graines ou fruits pour le réensemencement des champs et jardins afin d'assurer une récolte chaque année. A travers ce processus s'est opéré peu à peu une sélection (consciente ou inconsciente) d'individus présentant des phénotypes particuliers par les cultivateurs ce qui a conduit à une évolution anthropopisée de certaines espèces avec l'apparition de traits phénotypiques appelés « syndromes de domestication ». Selon l'étude phylogénétique de Salman-Minkov 2016 fondée sur l'analyse comparée de 107 genres et 2836 espèces d'angiospermes avec pour paramètre le nombre de chromosomes haploïdes de chaque génome étudié, il a été observé un enrichissement d'espèces polyploïdes pour les espèces domestiquées. Ainsi, 30% des espèces domestiquées présenteraient un état polyploïde dans leur phylogénie contre 24% pour les espèces relatives sauvages. Ce chiffre s'élève à 54% pour les espèces monocotylédones, de façon logique puisque la fréquence de polyploïdisation est beaucoup plus importante dans ce clade (41%) comparé au phylum des eudicotylédones par exemple (18%). La domestication d'espèces polyploïdes semble dépendre de l'organe utilisé pour la consommation humaine.

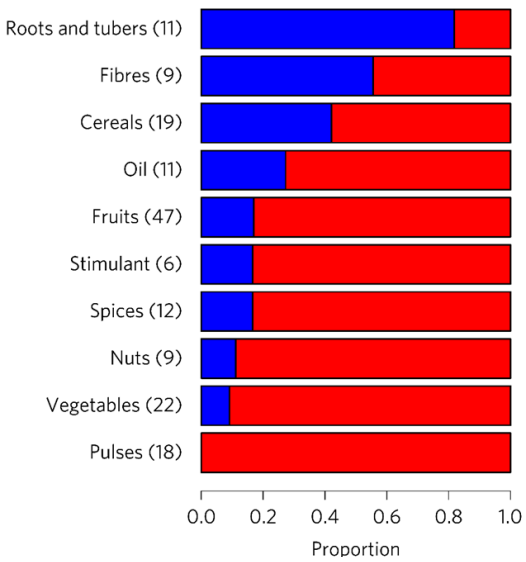


Figure 14. Proportions d'espèces polyploïdes en fonction de l'organe utilisé pour la consommation humaine.

Les nombres entre parenthèse correspondent aux nombres d'espèces dans chaque catégorie. Les couleurs bleu et rouge correspondent aux fractions d'espèces polyploïdes et diploïdes respectivement. Salman-Minkov 2016.

Une très forte proportion d'espèces polyploïdes domestiquées correspond à des plantes utilisées pour leurs tubercules et leurs racines ou encore pour leurs grains (céréales) (Figure 14). En effet, les syndromes phénotypiques associés à la domestication d'une espèce végétale correspond fréquemment à une augmentation de la taille de la partie consommée de la plante, une diminution de l'égrainage des grains chez les céréales, le passage d'un cycle de vie pérenne ou semi-pérenne à un cycle de vie annuel (mais, coton…) ou encore une forte adaptabilité à différents milieux de culture (*Triticum, Tehobroma, Coffea*). Il est d'ailleurs toujours débattu, pour un certain nombre d'espèces, de savoir si la polyploïdisation s'est produite avant leur domestication et a été sélectionnée et maintenue via ce processus ou bien si la domestication a elle-même favorisé l'apparition d'individus polyploïdes. Ceci est notamment le cas pour le blé tendre pour lequel certaines études archéo-

botaniques démontrent une collecte simultanée et systématique d'espèces polyploïdes (amidonnier tétraploïde par exemple) et d'espèces diploïdes du genre *Aegilops* par exemple (Kilian *et al.* 2010). Ainsi, la probabilité d'hybridation entre les individus provenant des graines de différentes espèces ressemées dans les champs de l'époque pouvait être largement augmentée.

Des études génétiques sur le blé et sur le coton ont démontré l'influence de la domestication sur certains traits via une évolution de certains *loci*. Chez le blé, Zhang *et al*. 2011 ont notamment démontré que les trois homéologues (A, B, D) du locus Q, responsable de l'adhérence des grains aux glumes (égrénage) et de la rigidité du rachis, avaient subi une hyper-fonctionnalisation (homéologue A), une pseudogénisation (homéologue B) et une sous-fonctionnalisation (homéologue D) au cours du processus de domestication. L'apparition de ce caractère a nécessité la mise en relation des trois gènes au sein du génome polyploïde avec mise en place d'une co-régulation spécifique et sélectionnée. Chez le coton, Zhang *et al*. 2015 ont quant à eux mis en évidence que le sous-génome A a été particulièrement impacté par une sélection anthropique biaisée pour les caractères liés à la qualité des fibres et que le génome D, lui, a plutôt subi une sélection pour des gènes impliqués dans la tolérance aux stress.

Ainsi, l'histoire évolutive des céréales polyploïdes cultivées, ainsi que celles d'un grand nombre d'autres espèces agronomiques polyploïdes, passe par le processus de domestication de ces espèces déjà polyploïdes ou en cours de polyploïdisation et dont l'évolution génomique a été influencée par ce processus. Le blé tendre constitue en ce sens une espèce modèle pour l'étude du comportement des gènes homéologues à la suite de la polyploïdisation.

II. Le blé tendre : un élégant système génétique pour étudier la polyploïdisation

Rubisko, Cellule, Apache, Arezzo, Boregar, Fructidor, sont des variétés de blé tendre (espèce *Triticum aestivum*) bien connues des cultivateurs français puisqu'elles représentent les cinq variétés les plus cultivées de la sole française (depuis 2010, France AgriMer). Ce blé dit tendre car panifiable, est peu à peu devenu une vedette des espèces cultivées et étudiées scientifiquement, indissociablement liée au destin de l'humanité pour deux raisons principales. La première est économique car cette espèce représente une des principales productions agricoles au niveau mondial (en compétition avec le Riz et le Maïs) avec un peu plus de 750 millions de tonnes produites en 2018, principalement par l'union Européenne, la Chine, l'Inde, la Fédération de Russie, les Etats-Unis et le Canada et 176 millions de tonnes en transactions internationales via la bourse de Chicago (FAO, Food Outlook, 2018). Grâce à ses grains riches de 13% de protéines (prolamines, gluténines, glutamines), le blé tendre permet de couvrir environ 21 % des besoins quotidiens en protéines des principaux pays producteurs (précités) et importateurs de blé (Egypte, Indonésie, Algérie, Brésil, Bangladesh, Japon, Phillipines, Mexique, Negeria) dont la consommation est fortement associée à l'acquisition du mode de vie occidental des populations (Henchion *et al*. 2017, Shewry et Hey 2015). La deuxième raison est scientifique car cette espèce présente un génome très

compliqué, fascinant pour découvrir de nouvelles caractéristiques biologiques des génomes eucaryotes ainsi qu'une grande diversité spécifique, témoins de la plasticité adaptative et des particularités évolutives des plantes.

II.1 Origine évolutive et géographique des Triticeae

D'un point de vue botanique, le blé et les espèces apparentées font partie de l'une des familles botaniques les plus riches en espèces : les Poaceae (10 000 et 12 000 espèces). La radiation évolutive des poaceae a été estimée vers la fin du Crétacé, il y a 55-70 MYA (Kellogg *et al*. 2001). Elle comprend neuf sous-familles parmi lesquelles les bambusoidées, panicoïdées (maïs, canne à sucre et sorgho), les Oryzoidées/Erhartoidées (riz) et les poïdées (blé, seigle, orge et avoine) (Hodkinson 2018).

Les relations simplifiées d'apparentement des espèces cultivées de Poaceae sont présentées sur la Figure 15 ci-après.

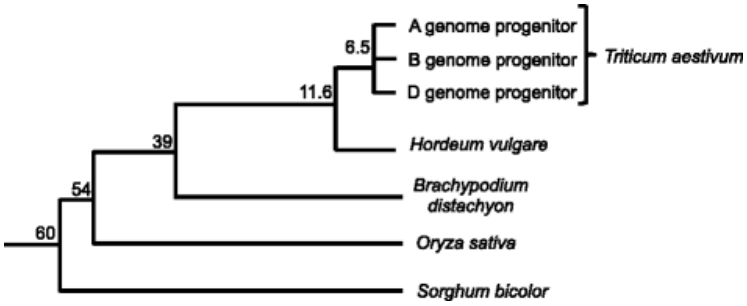


Figure 15. Phylogénie des espèces modèles de poaceae utilisée en génomique.

Glover et al. 2015

La tribu des Triticeae comprend 330 espèces qu'il est très difficile de distinguer morphologiquement entres elles tant les évènements d'hybridations interspécifiques voire inter-genres (*Triticum*, *Aegilops* et *Amblyopyrum*) sont courants au sein de cette tribu et favorisent les flux de gènes (Feldman and Levy 2015). Cette potentialité d'hybridation et d'allopolyploïdization a été révélée par de nombreuses études de cytogénétique qui ont mis en évidence les différents niveaux de ploïdie rencontrés dans cette tribu botanique : 31% de diploïdes (n=2, 7*2=14 chromosomes), 1% de triploïdes (n=3 , 7-3=21 chromosomes), 45% de tetraploïdes (n=4, 7*4=28 chromosomes), 17% d'hexaploïdes (n=6, 7*6=42 chromosomes) et les 6% restant allant de l'octoploïdie à la dodecaploïdie (Feldman and Levy 2015).

Ainsi, l'histoire des Triticeae semble être façonnée par la spéciation d'espèces diploïdes et l'apparition spontanée d'espèces polyploïdes. La divergence des espèces diploïdes, majoritairement autogames et dont le contenu en gènes est très conservé semble se faire principalement par l'évolution progressive ou saltatoire du contenu en TE (remobilisation inhérente au fonctionnement du génome ou bien liée à un stress environnemental, Senerchia *et al.* 2013). Les espèces polyploïdes issues d'une hybridation interspécifique (allopolyploïdes) sont quant à elles isolées d'un point de vue génétique et reproductif de façon soudaine, créant ainsi un saut évolutif. C'est exactement ce que raconte l'histoire évolutive du blé tendre *Triticum aestivum*. Le genre *Triticum* comprend seulement 6 espèces avec trois niveaux de ploïdie et quatre génomes distincts (Tableau 1).

Tableau 1. Tableau récapitulant les niveaux de ploïdie et les génomes composant le genre Triticum.

Espèce	**Génomes**	**Ploïdie et nombre de chromsomes**
Triticum monococcum L.	AA	2n=2x=14, n=7
Triticum urartu Tumanian	AABB	2n=2x=14, n=7
Triticum turgidum L.	AABB	2n=4x=28, n=14
Triticum timopheevii Zhuk	AABB	2n=4x=28, n=14
Triticum aestivum L	AABBDD	2n=6x=42, n=21
Triticum zhukovskyi	AAAAGG	2n=6x=42, n=21

L'histoire évolutive du genre *Triticum* est donc expliquée par des événements d'hybridation et de polyploïdisation inédits et est indéniablement reliée à l'histoire de l'agriculture qui a participé à sa diversification. Aujourd'hui, le blé tendre hexaploïde compte plus de 150 000 accessions provenant d'une centaine de pays (CYMMIT). L'analyse phylogéographique de 4506 accessions a permis de comprendre l'histoire de la dispersion de cette céréale dans le monde (Balfourier *et al*. 2019). Les auteurs de l'étude ont mis en évidence une chronologie des pôles de divergence à partir de son apparition dans la région du croissant fertile, avec un pôle associé à la région du bassin méditerranéen à partir duquel deux voies de diversification se sont développées : l'une empruntant la route du Danube vers l'Europe et l'autre foulant la route de la soie vers l'Asie (Figure 33).

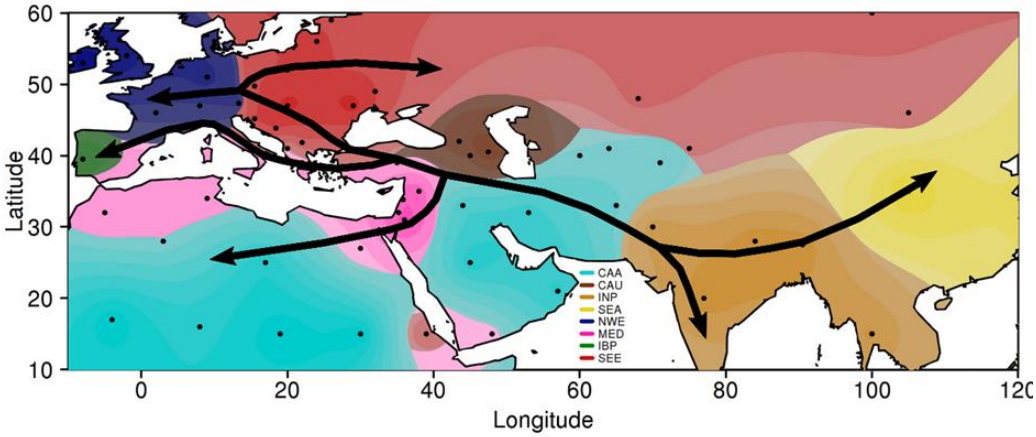


Figure 16 Carte retraçant la dispersion des landraces de blé à travers le monde.
Carte établie en utilisant des données génomiques de 4506 landraces et cultivars provenant de 105 pays (génotypage SNP notamment), Balfourier et al. 2019.

II.2 Histoire évolutive de l'espèce *Triticum aestivum*

A travers l'étude des données génomiques produites pour la publication d'une première ébauche de séquence génomique de référence du génome de *Triticum aestivum* (accession Chinese Spring, IWGSC en 2014), les analyses de Marcussen et ses collaborateurs (2014), a conforté l'histoire évolutive de cette espèce. Grâce à l'analyse d'arbres phylogénétiques de gènes et des espèces appartenant à la tribu des Triticeae (*Triticum aestivum, T. monococcum, T. urartu, Ae. sharonensis, Ae. speltoides, and Ae. Tauschii),* les auteurs ont confirmé et daté la double hybridation interspécifique ayant conduit au génome du blé tendre. La première hybridation s'est déroulée entre les espèces *Aegilops speltoïdes* (génome BB) et *Triticum urartu (génome AA)* et a donné l'espèce tétraploïde *Triticum turgidum sp. Diccocoïdes (génome AABB)* il y a environ 800 000 ans. La deuxième aurait eu lieu entre *Triticum turgidum* sp. *Diccocoïdes* et *Ae. tauschii* il y a environ 10 000 ans (Figure 17). Ils ont également mis en évidence que l'espèce *Aegilops tauschii* portant le génome D dériverait elle-même d'une hybridation homoploïde (sans changement du niveau de ploïdie) entre les espèces *Aegilops speltoïdes* (génome BB) et *Triticum urartu (génome AA)* il y a environ 5.5 millions d'années (2,5-3MA selon les analyses sur ADN chloroplastique de Middelton *et al.* 2014). La dernière hybridation interspécifique donnant le blé hexaploïde serait concomitante à la phase de sédentarisation et de développement de l'agriculture dans les plaines du croissant fertile et d'Anatolie. Des études archeo-botaniques ont identifié que ce serait sur les bords de la mer Capsienne où des populations de blé tétraploïdes alors en cours de domestication et des populations sauvages d'espèces diploïdes *Ae. tauschii* se seraient hybridées, conduisant à l'apparition de l'hexaploïde *Triticum aestivum* (Salamini *et al*. 2002, Matsuoka 2011). La domestication de différentes espèces de *Triticum* apparentées ayant différents niveaux de ploïdie au sein d'un même bassin de développement de l'agriculture a également favorisé des transferts de gènes par de probables hybridations supplémentaires entre les différentes espèces en cours de domestication : *Triticum monoccocum* (engrain ou petit épeautre, AA, 2n=14), *Triticum turgidum spp. diccocoïdes* (amidonnier sauvage, AABB 2n=28), *Triticum turgidum spp. diccocon* (amidonnier AABB 2n=28), *Triticum aestivum spp. spelta* (grand épeautre ou blé des gaulois AABBDD 2n=42) et *Triticum aestivum spp. aestivum* (froment ou blé tendre AABBDD 2n=42) (Feldman et Levy 2015, Pont *et al.* 2019).

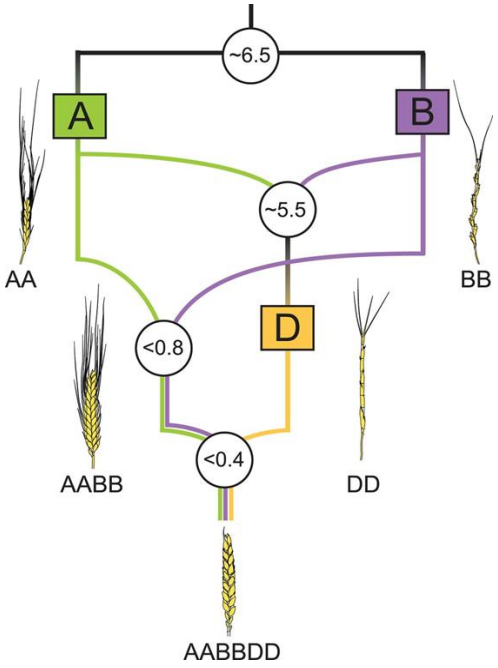


Figure 17. Modèle de l'histoire évolutive du blé tendre.

Les chiffres dans les cercles blancs correspondent aux dates approximatives des hybridations interspécifiques et des divergences en millions d'années. Marcussen et al. 2014.

II.3 Composition du génome de blé tendre cv Chinese spring, première accession de blé tendre séquencée

II.3.1 Séquence de référence du génome du blé tendre

Le séquençage du génome du blé tendre a été un véritable défi scientifique tant sa complexité en termes de taille (15Gb) et de taux de séquences répétées (85%) était importante. En effet, par rapport à d'autres espèces à grands génomes (*Pinus taea*, 22Gb, plus grand génome séquencé) ou à génomes polyploïdes (Colza, ou coton avec 1.4 et 2.5 Gb respectivement), le génome du blé combine simultanément ces deux contraintes majeures : taille et proportion de TEs importantes. C'est en 2018 qu'a été publiée la première séquence de référence complète et de haute qualité du génome de l'accession *Triticum aestivum* cv Chinese Spring (IWGSC, 2018) grâce à l'algorithme d'assemblage développé par la société NRGene (DeNovoMagic®). Cette séquence représente 14.5 Gb pour 21 pseudomolécules (chromosomes), soit 94 % de la taille de génome attendue avec une N50 pour les scaffolds de 7Mb. Chacune des pseudomolécules des chromosomes de ce génome a été reconstruite en utilisant en moyenne 76 superscaffolds. En 2014, la construction de la pseudomolécule du chromosome 3B avait nécessité 2808 scaffolds présentant une N50 de 892 kb soit 8 fois plus faible que celle de la séquence génomique complète de 2018. La part des scaffolds n'ayant pu être ancrés sur les pseudomolécules représente seulement 2.8 % de la taille totale de l'assemblage (avec seulement 3 % de séquences géniques) ce qui confirme également la qualité de cette séquence de référence.

II.3.2 Description du contenu en gènes

II.3.2.1 Annotation des gènes

L'annotation de cette séquence de référence pour identifier les gènes a été réalisée conjointement entre le GDEC (avec l'outil d'annotation TriAnnot (Leroy *et al.* 2012)), l'Earlham Institute (Norwich, UK) et l'IPK (Gatersleben, Allemagne) via l'utilisation de deux « pipelines » d'annotation différents (Triannot et PGSB) et a permis l'identification de 107 981 gènes, dits de haute confiance « Hight confidence » (gènes HC) codant pour des protéines avec pour chaque sous-génomes : A : 35 345 ; B : 35 643 et D : 34 212 gènes. Ces gènes ont notamment été validés par des données d'expression pour 94,114 (84.9%) d'entre eux et par une prédiction de fonctions pour 82.1% (90,919). En plus de ces gènes, 161 537 séquences putatives de gènes dits de faible confiance « low confidence » (gènes LC) ont été identifiées, correspondant à des fragments de gènes. Pour cet ensemble de gènes, seuls 49% d'entre eux présentaient des données d'expression pouvant indiquer une réelle fonction. Enfin, 303,818 pseudogènes ont été identifiés avec 8% d'entre eux inclus dans la catégorie des gènes LC. La comparaison des sous-génomes en termes de pseudogènes a révélé une proportion significativement plus faible de ces gènes présents sur le sous-génome D. La qualité de l'annotation a été validée par la recherche de la proportion de gènes orthologues du blé tendre correspondant aux 1440 gènes modèles universaux des embryophytes (méthode BUSCO, Simão *et al.* 2015). Avec une proportion de 99% de gènes modèles retrouvés en utilisant au

moins une copie homéologue et 90% retrouvés en utilisant les trois copies, l'annotation a été considérée comme étant de qualité.

II.3.2.2 **Inférence des gènes homéologues chez le blé tendre**

Afin de caractériser les relations d'homéologie entres les gènes des trois sous-génomes du blé tendre, une analyse phylogénétique a été réalisée avec les séquences des espèces diploïdes (*Triticum urartu, Aegilops tauschii*), d'autres espèces de poaceae (*Oryza sativa, Zea mays, Sorghum bicolor, Brachipodium dystachion, Hordeum vulgare, Secale cereale*) et de plantes plus éloignées phylogénétiquement (viridiplantae *Arabidopsis thaliana, Physcomitrella patens, Selaginella moellendorffii, Chlamydomonas reinhardtii).* Cette analyse a été menée sur un ensemble de gènes comprenant 181,036 gènes avec 103 757 gènes « Hight confidence » et 77 279 gènes « low confidence ». 39 238 groupes de gènes homéologues ont été identifiés à travers l'analyse conjointe des arbres de gènes comportant 63% des 181 036 gènes (IWGSC 2018, Tableau 2).

Tableau 2. Tableau présentant les différents groupes d'homoéologie au sein du génome du blé tendre (IWGSC 2018)

Homeologous group (A:B:D)	Number in wheat genome	Composition of groups (%)	Number of genes in A	Number of genes in B	Number of genes in D	Total number of genes
1:1:1	21,603	55.1	21,603	21,603	21,603	64,809
1:1:N	644	1.6	644	644	1,482	2,770
1:N:1	998	2.5	998	2,396	998	4,392
N:1:1	761	1.9	1,752	761	761	3,274
1:1:0	3,708	9.5	3,708	3,708	0	7,416
1:0:1	4,057	10.3	4,057	0	4,057	8,114
0:1:1	4,197	10.7	0	4,197	4,197	8,394
Other ratios	3,270	8.3	4,999	5,371	4,114	14,484
1:1:1 in microsynteny	18,595	47.4	18,595	18,595	18,595	55,785
Total in microsynteny	30,339	77.3	27,240	27,063	28,005	82,308
1:1:1 in macrosynteny	19,701	50.2	19,701	19,701	19,701	59,103
Total in macrosynteny	32,591	83.1	29,064	30,615	30,553	90,232
Total in homeologous groups	**39,238**	**100.0**	**37,761**	**38,680**	**37,212**	**113,653**
Conserved subgenome orphans			12,412	12,987	10,844	36,243
Nonconserved subgenome singletons			10,084	12,185	8,679	30,948
Nonconserved subgenome duplicated orphans			71	83	38	192
Total (filtered)			**60,328**	**63,935**	**56,773**	**181,036**

Parmi ces groupes, 55% sont des triplets d'homéologues 1:1:1, c'est-à-dire une copie de gènes homéologues identifiée sur chaque chromosome homéologue, (par exemple 1A-1B-1D). Ainsi, presque la moitié des groupes d'homéologues ne correspond finalement pas des triplets. Notamment, 15% des groupes présente des gènes dits in-paralogues (issus de duplications non liées à la WGD) avec des nombre de copies supplémentaires variant de 2 (exemple 2 :1 :1) à 14 (exemple 14:2:0). Pour l'ensemble des groupes identifiés, le pourcentage global de gènes affilié à chaque sous génome est équilibré 63%(A), 61%(B), et 66%(D) indiquant qu'il n'y a pas de fractionnement biaisé ni de duplications de séquences préférentiellement sur l'un des trois sous-génomes. En effet, les pourcentages de groupes n'ayant que deux copies homéologues sont similaires quel que soit le sous génome présentant cette absence : 10.7% (0:1:1) sous-génome A, 10.3% (1:0:1) sous génome B et 9,5% (1:1:0) sous-génome D. En revanche, il a été observé 1) une perte progressive de certains gènes (78 familles de gènes en contraction) et 2) des mouvements de gènes entre les sous-génomes qui se seraient produits lors (i) de la spéciation des espèces progénitrices ou bien (ii) à la suite à l'hybridation entraînant la formation du tétraploïde.

Parmi les 181,036, 37% ne sont pas affiliés aux groupes d'homoéologie présentés précédemment et 46% de ces 67,383 gènes ne présentent pas non plus d'orthologues ni dans les espèces apparentées au blé ni dans les espèces plus éloignées évolutivement. Parmi ces gènes, ont été identifiés le gène de la synthèse de granules d'amidon (granule bound starch synthase, GBSS) sur le chromosome 4A (1 :0 :0) et le gène ZIP4 présent dans le locus Ph1 sur le chromosome 5B (0:1:0), impliqué dans le comportement de méiose de type diploïde (chromosome bivalents) des chromosomes homologues. Ces gènes uniques à ce génome ont donc probablement fortement participé à l'évolution de cette espèce.

Une autre caractéristique importante de ce génome en termes de contenu génique est la présence d'une proportion importante de gènes dupliqués en tandem comprenant 27% des gènes HC, proportion 10% plus élevée que celle retrouvée chez d'autres espèces de monocotylédones. Un léger biais envers le sous génome B a été détecté avec une proportion plus importante de gènes appartenant au groupe 1 :N :1. Les auteurs ont également identifié des différences en termes de divergence de séquence entre les groupes de gènes homéologues comportant des gènes dupliqués (N :1 :1, 1 :N :1 et 1 :1 :N) et les gènes des groupes 1 :1 :1 et 1 :1. Les premiers présentent des taux de mutations synonymes (ks) plus élevés par rapport aux seconds, pour le sous génome présentant les copies dupliquées. Ceci peut traduire un relâchement de la pression de sélection favorable à l'innovation évolutive pour ces gènes paralogues.

II.3.2.3 **Structure fonction et évolution du chromosome 3B**

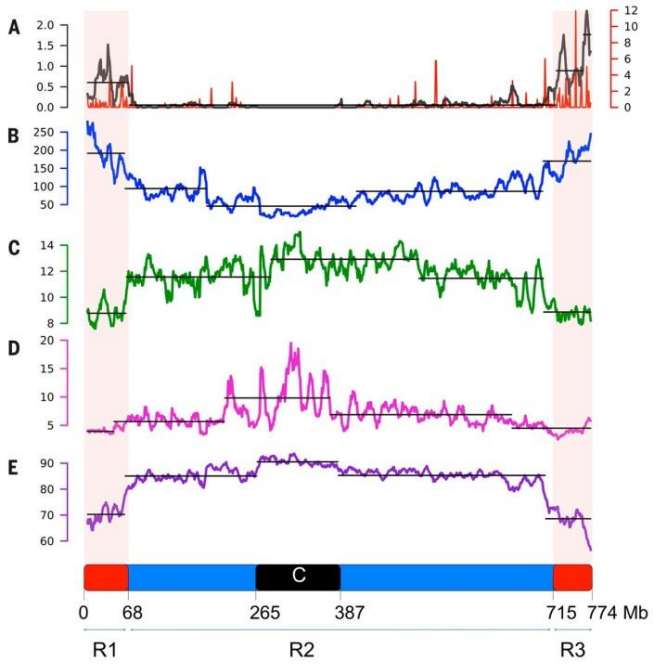


Figure 18. Partitionnement structural et fonctionnel du chromosome 3B du blé tendre.

A)Taux de recombinaison méiotique (en cM/Mb) dans une fenêtre glissante de 10mb en noir et 1Mb en rouge ; B) Densité de gènes (en CoDing Sequence CDS/10Mb) ; C) Amplitude d'expression des gènes (15 conditions/tissus/ stades de développement RNA-seq) ; D) Moyenne du nombre de transcrits alternatifs par gènes ; E) Densité d'Eléments transposables ; R1 et R3 = régions distales, en bleu = régions péricentromériques, C = centromère. Choulet et al. 2014.

L'équipe Seven (Structure et Evolution du Génome du Blé), au sein de laquelle j'ai effectué mon doctorat, a été l'un des piliers de l'IWGSC depuis sa fondation. Avec le chromosome 3B comme modèle d'étude, l'équipe a produit la première carte physique (Paux *et al.*, 2008), les premières grandes séquences de plusieurs Mb (Choulet *et al.*, 2010) et a publié en 2014 la première pseudomolécule de haute qualité (~774 Mb, N50 : 892 kb) pour le chromosome 3B (Choulet *et al.*, 2014) (Figure 18). Cette étude a permis de déterminer avec précision les caractéristiques structurales et fonctionnelles d'un chromosome en particulier, servant de base à l'analyse de l'ensemble du génome. Les variations observées des taux de recombinaison et densité en gènes le long du chromosome ont permis de définir cinq régions chromosomiques aux caractéristiques bien distinctes. Les régions terminales R1 et R3 présentent une proportion plus faible de TE, une

densité de gènes plus importante, en particulier en gènes non-synténiques et à expression non-constitutive ainsi qu'un taux de recombinaison plus élevé (Figure 18) comparé aux régions péricentromériques R2a et R2b. Une analyse détaillée des gènes dupliqués sur ce chromosome avait mis en évidence que 87% des gènes non-synténiques avec *Brachypodium*, le riz et le sorgho présentaient un paralogue sur un autre chromosome, pointant les duplications interchromosomiques comme un mécanisme majeur à l'origine de ces gènes non- synténiques. Par ailleurs, 46% de ces gènes sont dupliqués en tandem.

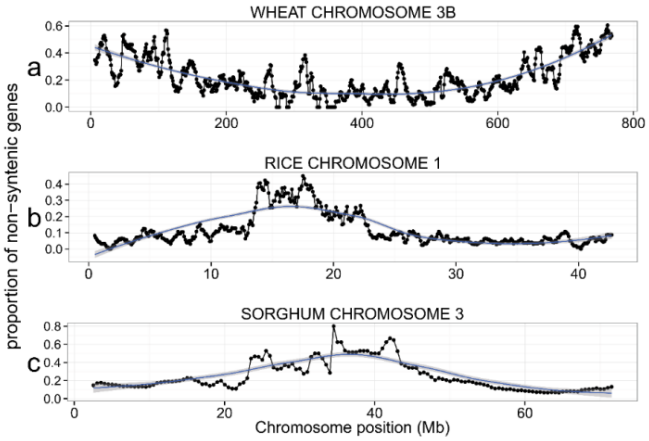


Figure 19. Distributions chromosomiques des gènes non-synténiques chez trois espèces de poaceae.

Cette figure représente la distribution des proportions des gènes non synténiques sur le total de gènes dans la fenêtre considérée pour a) le chromosome 3B du blé, b) le chromosome 1 du riz et c) le chromosome 3 du sorgho. Pour chaque espèce, la fenêtre glissante d'analyse était de 10Mb, pas 1MB, 1Mb pas 0,1Mb et 5Mb pas 0,5MB respectivement. Les pseudogènes ont été enlevés de l'analyse. Glover et al. 2015.

L'étude pionnière du chr3B avait aussi abouti à une conclusion nouvelle qui était l'importance considérable des duplications de gènes sur l'évolution du contenu en gènes des *Triticeae* par rapport aux autres *Poaceae*. En effet, dans les 27% de gènes issus de duplications récentes (après la divergence avec *Brachypodium*), 1000 gènes non-synténiques (sur ~7000), uniques au génome du blé tendre ont été retrouvés sur le chr3B, particulièrement abondants dans les régions distales (Figure 19, Glover *et al.* 2015).

II.3.3 **Polyploïdisation et régulation du transcriptome chez le blé tendre**

II.3.3.1 **Avant l'obtention de la séquence génomique de référence complète**

Chez le blé, l'obtention récente de la séquence génomique de référence sur l'ensemble des espèces progénitrices, polyploïdes et celles d'individus synthétiques ou hybrides F1 n'a permis de constituer des analyses transcriptomiques sur l'ensemble des gènes et du corpus d'espèces que très récement.

Avant cela, une étude transcriptomique sur puce à ADN pour une accession de blé hexaploïde resynthétisée avait démontré que sur 34 000 gènes homéologues identifiables, 19% d'entre eux présentaient une expression non-additive (le niveau d'expression du polyploïde ne se situaient pas entre les niveaux des espèces progénitrices) (Akhunova *et al*. 2010). Ce biais d'expression correspondait à une activité transcriptionnelle de la copie du parent dominant plus importante (seuls 6 à 11% de ces biais d'expression correspondaient à des diminutions d'expression des sous-génomes relatifs). Cette étude avait cependant démontré que 84% de ces expressions non-additives pouvaient être expliquées par la différence entre les génotypes utilisés pour recréer la variété synthétique. D'autres études sur des blés hexaploïdes synthétiques basées sur des données de puces Affimetrix telles que celle de Chagué *et al*. 2010 et Chelaifa *et al*. 2013 avaient mis en évidence une prépondérance de l'expression additive des gènes (« mid parent

values expression »). Ces analyses confortaient donc l'hypothèse d'une absence de remaniement massif de l'expression des gènes homoelogues au sein du génome polyploïde du blé tendre.

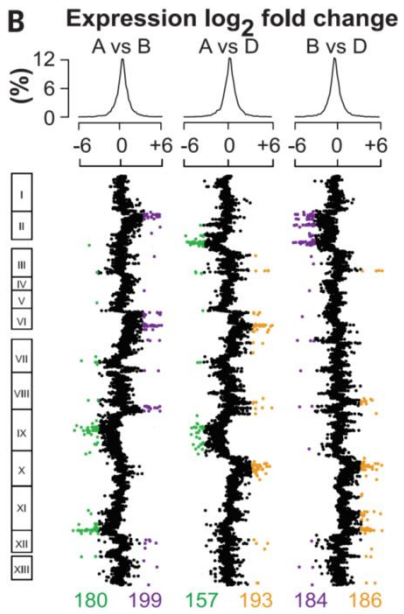


Figure 20. Comparaison des niveaux d'expression des paires de gènes homéologues en fonction des groupes d'expression dans différents organes (chiffres romains à gauche).

Les points en couleurs correspondent aux paires de gènes présentant un différentiel d'expression significatif (p-value < 0,05) entre les deux homéologues : vert = sous-génome A, violet = sous-génome B et orange = sous-génome D. IWGSC 2014

En 2014, en parallèle de l'obtention de la pré-séquence de référence, des données d'expression RNA-seq avaient été produites et permis de comparer l'expression relative des copies homéologues à l'échelle du génome entier. L'analyse de données transcriptomiques de 6 219 triades (gènes présents en trois copies homéologues 1 :1 :1) sur cinq organes différents (épi, grains, racines feuilles, tiges) avait révélé que chaque sous-génome présente un haut degré d'autonomie dans l'expression des copies. La comparaison des données d'expression par paires d'homéologues A vs B, A vs D et B vs D ne montrait un biais d'expression des copies homéologues que pour 21% des paires étudiées. De façon globale aucune dominance de l'un des sous-génomes en termes de transcription n'avait été révélée par ces résultats, illustrés dans la figure 20 ci-contre. De plus, cette absence de dominance concernant l'expression des copies homéologues était associée à l'absence d'un fractionnement biaisé de l'un des sous-génomes. Dans leur analyse de la séquence publiée, aucune perte massive de séquences sur l'un des sous-génomes n'avait été identifiée. Une perte des copies provenant de familles de gènes en expansion avait été observée, mais indépendante des sous-génomes. Seule une légère différence en termes de nombre de copies était observée avec moins de copies perdues sur le sous-génome D. Cette différence était expliquée par le fait que ce sous-génome évoluant dans le génome polyploïde depuis moins longtemps que les sous-génomes A et B (hybridation datée de 10 000 vs 800 000ans), il aurait été moins impacté par les remaniements liés au phénomène de diploïdisation post polyploïdisation.

En parallèle de cet article, Pfeifer *et al.* 2014 ont étudié l'expression des gènes homéologues dans trois types cellulaires (couche aleurone, les cellules de transfert et l'endosperme) pour trois stades de développement du grain (10, 20 et 30 jours après anthèses). Une analyse en composante principale avait regroupé prioritairement les données d'expression par sous-génome et non par tissu ou stade de développement. Néanmoins, une analyse détaillée des données d'expression en regroupant les gènes selon des profils d'expression spécifiques au sein des assises cellulaires et au cours des différents stades de développement avait permis de mettre en évidence la dominance de l'un des trois sous-génomes alternativement au sein des 23 profils d'expression définis mais sans démontrer la dominance globale d'expression de l'un des sous-génomes sur les deux autres. Ils avaient également analysé les profils

d'expression des différentes familles de gènes caractéristiques du fonctionnement du grain (comme les protéines de réserve : gluténine, gliadines…) et avaient montré que certains gènes de l'un des sous-génomes présentaient une dominance d'expression. Par exemple, l'expression globale des gènes codant pour les gluténines de faible poids moléculaire provenait à 68% des copies du génome B. Enfin, les auteurs avaient montré l'activation alternative de certains domaines chromosomiques pour chacun des sous-génomes pour un tissu ou un stade de développement donné, indiquant une dominance aléatoire des sous-génomes en fonction des stades de développement.

L'étude de Leach *et al*. 2014 sur les gènes homéologues des chromosomes 1 et 5 avait montré une expression biaisée pour 26% des homéologues. Cependant, ces travaux n'avaient ciblé spécifiquement que les gènes homéologues en triplet. De même, l'étude de Harper *et al*. 2016 n'avait démontré aucune dominance d'expression de l'un des sous-génomes pour 15 527 triplets de gènes et que certaines régions chromosomiques seraient préférentiellement associées à des biais d'expression entre les trois homéologues.

Une analyse de transcriptomique sur un blé synthétique et ses progéniteurs pour trois tissus différents (jeune plantule, jeune épi, et jeune grain) avait montré que les gènes présentant une expression non-additive étaient plutôt rare mais que les profils de non-additivité sont retrouvés jusqu'à la quatrième (et dernière) génération étudiée (Li *et al*. 2014). Ils ont aussi montré une EDL (Expression Level Dominance) en faveur des gènes AABB sur les gènes DD, mais dans leur étude, la proportion de gènes dans ce cas de figure dépendait du tissu étudié. Ils avaient noté à ce sujet que les tissus présentaient des tailles et potentiellement des niveaux de transcription sensiblement différents entre les espèces progénitrices et polyploïdes.

Ces première études, réalisées presque uniquement sur les homéologues en triplets ont révélé une autonomie dans l'expression des trois sous-génomes avec des dominances spécifiques de certains groupes de gènes de l'un des sous-génomes selon les tissus ou stades de développement considérés. D'autres ont mis en évidence des biais d'expression entre homéologues et des pertes de séquences plus marquées pour les sous génomes A et B comparativement au sous-génome D.

II.3.3.2 **Après l'obtention de la séquence génomique de référence**

En 2018, la publication de la séquence de référence de l'accession Chinese spring a été accompagnée de la publication d'un atlas transcriptomique compilant 850 échantillons de RNA-seq provenant de 32 tissus prélevés dans des conditions différentes (stress biotiques et abiotiques, contrôle). Une expression significative (expression supérieure à 0 ,5TPM dans au moins un des tissus, moyenne de trois réplicas biologiques) a été observée pour 85% des gènes HC et pour 49% des gènes LC. En moyenne, les gènes HC présentent des niveaux d'expression de 8,2 tpm et une amplitude d'expression de 20 tissus contre 2,9 tmp et 6 tissus pour les gènes LC. 8 321 gènes HC ont été identifiés comme exprimés spécifiquement dans un tissu (expression tissus spécifique) et 23 146 sont exprimés dans les 32 tissus. A l'échelle du

génome, les gènes situés dans les régions distales présentaient plus fréquemment une expression plus faible et plus spécifique en comparaison des gènes présents dans les régions proximales. Grâce à l'annotation de qualité des gènes et l'affiliation de chacun à des groupes d'homéologie, une analyse concernant les biais d'expression des gènes homéologues en triplets (1 :1 :1) a été réalisée et est présentée dans ce manuscrit de thèse dans le chapitre 2.

III. Apport de l'épigénétique pour la compréhension de l'évolution d'un génome polyploïde

L'épigénétique est une discipline de la biologie visant à étudier les processus moléculaires transmissibles par mitose et/ou méiose modulant l'expression des gènes sans changement de la séquence nucléotidique (Johannes *et al.* 2008). Différentes marques épigénétiques vont être impliquées dans la structure de la chromatine, le recrutement de protéines de régulation et l'accessibilité aux machineries de transcription/réplication de l'ADN. Dans un génome polyploïde, l'hypothèse épigénétique qui est explorée actuellement réside dans l'analyse du différentiel de marquage épigénétique des séquences homoéologues pouvant entraîner l'expression différentielle des gènes.

III.1 Mécanismes épigénétiques étudiés

III.1.1 La méthylation de l'ADN

La méthylation de l'ADN est historiquement la première marque épigénétique à avoir été étudiée et est encore la plus étudiée actuellement de par la facilité de sa détection à l'échelle du génome, suffisamment résolutive pour identifier des épiallèles ou des régions différentiellement méthylées. Biochimiquement, chez les eucaryotes, elle correspond à la liaison d'un groupement méthyle en position 5 de l'anneau pyrimidique de la cytosine (5 mC) par une enzyme de type methyltransférase qui utilise la S-adénosylméthionine (SAM) comme donneur de groupement méthyle. Contrairement à la plupart des métazoaires qui présentent uniquement des méthylations sur des sites CG, chez les plantes, les cytosines sont méthylées selon trois contextes CG, CHG, CHH où H représente une Adénine, une Cytosine ou une Thymine (Henderson & Jacobsen, 2007, Figure 21). Cette marque épigénétique intervient dans la modulation de l'expression des gènes et dans la répression TE et ses patrons de localisation sont transmissibles à travers les générations. La mise en place de la méthylation de l'ADN chez les plantes se fait via une voie dépendante de fragments d'ARN qui permettront de cibler un locus particulier, voie appelée RdDM (RNA-Directed DNA méthylation). Une autre voie utilise une autre polymérase qui intervient dans la répression des éléments transposables transcriptionnellement actifs (McCue *et al.* 2015).

L'analyse des profils de méthylation le long des chromosomes montrent qu'il existe des régions nettement moins méthylées que d'autres et un enrichissement de contexte de méthylation selon le type de séquence ciblé (TE, gènes, pseudogènes) qui définiront des régulations différentielles de ces séquences, résumées dans la légende de la figure 21. Ainsi, au niveau des séquences géniques, la méthylation dans le corps des

gènes est associée à une transcription active alors que celle au niveau des régions promotrices est plutôt inhibitrice.

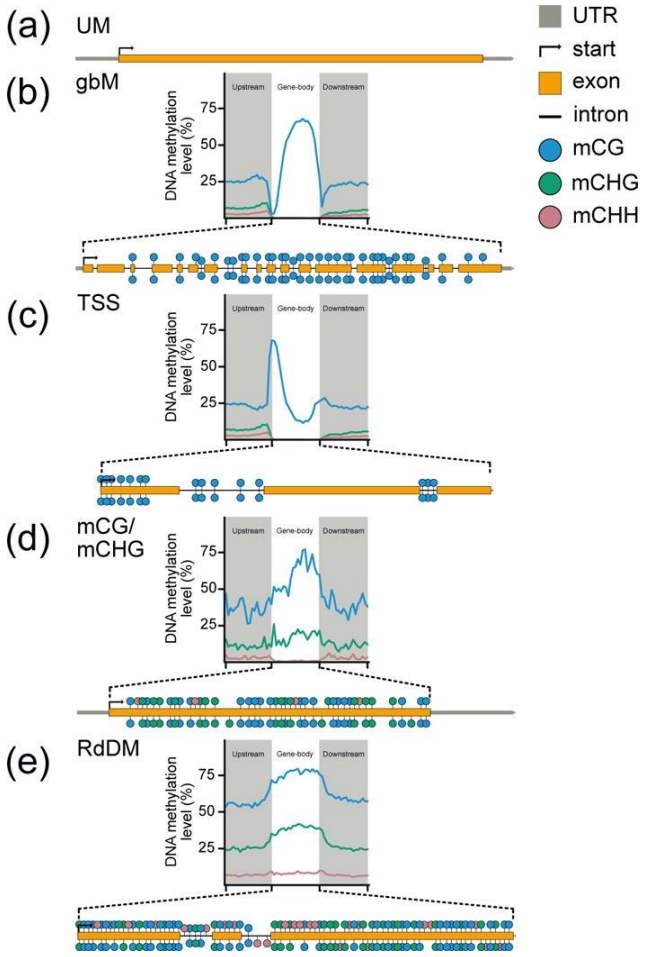


Figure 21. Profils schématiques de méthylation de l'ADN des gènes chez les plantes terrestres.

a) *Gènes non méthylés (Unmethylated : UM) ; absence de méthylation sur l'ensemble de la région transcrite du corps du gène pour les trois contextes CG, CHG et CHH*

b) *Gènes dont le corps est méthylé (Gene body methylated : gbM) ; enrichissement en méthylation pour le contexte CG de la région transcrite e perte de méthylation au niveau du site initiateur de la transcription (Transcription Start Site : TSS) et au niveau du site terminateur (Transcription Termination Site : TTS) ; gènes possédant une expression constitutive et ubiquitaire plus élevée que les gènes non gbM*

c) *Gènes méthylés au niveau de leur TSS ; enrichissement en méthylation de type CG uniquement au niveau du TSS du gène avec répression de son expression*

d) *Gènes dits CG/CHG ; enrichissement en méthylation de CHG et perte de méthylation de type CHH au niveau du corps du gène, avec la présence ou non de méthylation de type CG ; gènes exprimés à des niveaux très faibles par rapport aux autres gènes et aux gènes ayant une méthylation gbM chez les angiospermes (Niederhuth et al.2016).*

e) *Gènes dits CHH/RdDM ; enrichissement en méthylation CHH accompagné ou non de méthylation CG et ou CHG au niveau du corps du gène ; gènes réprimés dans l'ensemble de la plante chez A. thaliana à l'exception du pollen et des graines en développement. Bewick et Schmitz, 2017*

Dans le cadre de l'analyse des génomes polyploïdes, le phénomène de « spreading » (étalement) de la méthylation des TE vers les séquences adjacentes pourrait expliquer des différences d'expression entre gènes ou sous-génomes homéologues. En effet, la méthylation mise en place par la voie RdDM au niveau des TE peut se propager sur environ 300 pb avec des conséquences possibles sur les gènes flanquant ces TEs (Ahmed *et al.* 2011b). Ainsi, des proportions de TE différentes entre deux sous-génomes homéologues pourraient alors expliquer des différences d'expression des gènes via ce processus épigénétique (exemples décrits dans Mirouze and Vitte 2014, Do Kim *et al.* 2014).

III.1.2 **Les marques histones et les états chromatiniens**

La chromatine est une structure moléculaire composée de la molécule d'ADN et de nucléosomes permettant sa compaction dans le noyau des cellules eucaryotes. Les nucléosomes sont des complexes protéiques constitués d'un octamère de protéines appelées histones (H2A, H2b, H3, H4) autour duquel s'enroulent 147 paires de base de la molécule d'ADN (McGinty et Tan 2016). Les parties N-terminales des histones sont exposées à l'extérieur des nucléosomes et sont sujettes à des modifications post-traductionnelles de types méthylations, acétylations, ubiquitinations et phosphorylations… (Figure 22).

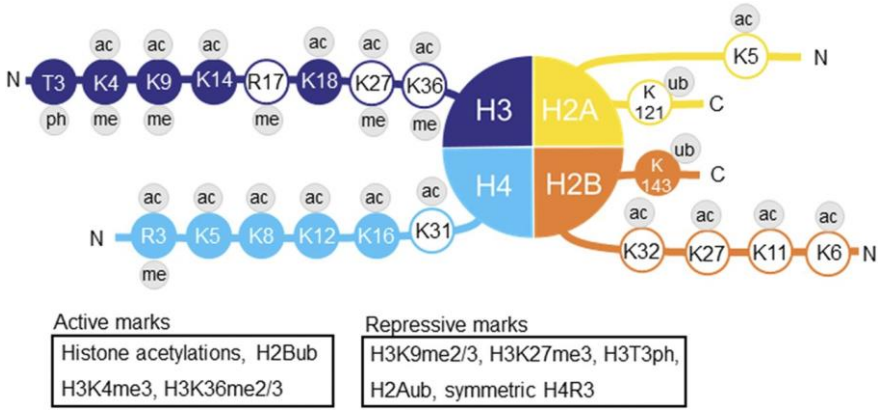


Figure 22. Représentation schématique d'un nucléosome et des différentes modifications post-traductionnelles des parties N-terminales des histones.

Ueda et Seki 2020

Selon la densité de nucléosomes et le type de groupements chimiques apposés sur les parties N-terminales, le différentiel de charges positives *versus* charges négatives en interaction avec la molécule d'ADN, qui elle est chargée négativement, va entraîner une compaction plus ou moins importante. En particulier, une présence accrue de groupements acétyles entraine une diminution des charges basiques (+) au niveau des histones et un relâchement des interactions nucléosomes/ADN. *A contrario*, les groupements méthyles favorisent l'interaction ADN/protéines en créant des sites d'accrochage de protéines se liant à l'ADN ce qui augmente la compaction (Udea et Seki 2020, Deal et Henikoff 2011). Ces protéines recrutées en fonction des combinaisons de groupements chimiques des histones vont modifier la forme, l'encombrement et l'enchevêtrement de la chromatine.

La densité en nucléosomes, les modifications chimiques des histones ainsi que les repliements spécifiques donnent une forme particulière à la fibre de chromatine qui joue un rôle important dans la transcription des gènes, la réplication, la réparation, ou encore la condensation des chromosomes lors de la mitose.

La connaissance de la distribution et de la dynamique de ces combinaisons de marques histones le long des chromosomes permet de définir des états chromatiniens aux caractéristiques fonctionnelles spécifiques.

III.1.3 **Définition et fonction de régulation des états chromatiniens**

III.1.3.1 **Les différents états de la chromatine**

La revue de Vergara et Gutierrez 2017 résume les travaux obtenus chez *Arabidopsis thaliana* et *Zea mays* entre autres pour décrire quelles marques histones sont associées aux différents états de la chromatine (Figure 23).

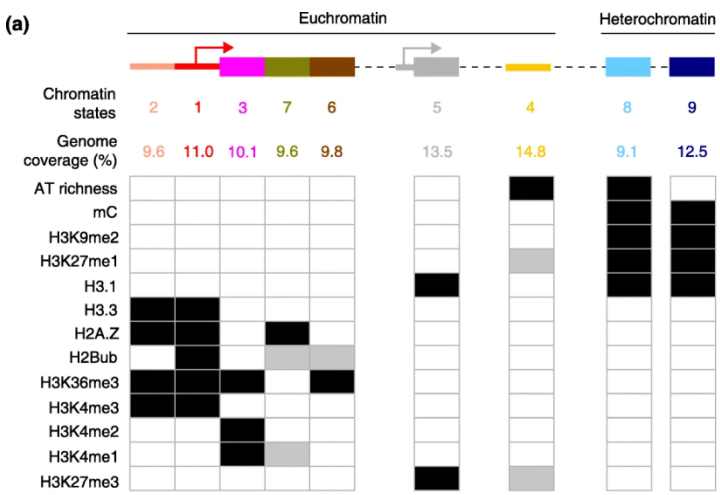


Figure 23. Caractérisation des principaux états chromatiniens présents dans un génome eucaryote (à partir d'études épigénomiques de Drosophila melanogaster, Ceanorabditis elegans, Zea mays).

Stade 2 = promoteur proximal
Stade 1 = Début du site de transcription
Stade 3 = partie 5' de la séquence codante
Stade 7 = longues séquences codantes
Stade 6 = partie 3' de la séquence codante
Stade 5 = chromatine "polycomb"
Stade 4 = regions intergéniques régulatrices distales
Stade 8 = hétérochromatine riche en bases AT
Stade 9 = hétérochromatine riche en bases CG
La densité des marques pour chaque état est représentée par les nuances de gris : carrés noirs = forte densité, carrés gris = densité moyenne et carrés blancs = densité très faible ou absence de marque.
Vergara et Gutierrez 2017

Pour résumer, les combinaisons de marques histones définissant les trois principaux états chromatiniens sont décrites ci-dessous :

L'hétérochromatine a été associée à des niveaux de méthylation des cytosines de l'ADN (5mC, en particulier en contexte CHH) élevés ainsi qu'à la présence accrue des marques histones H3K9me1 et H3K9me2 (Bernatavichute *et al*. 2008). Cet état chromatinien est cantonné au niveau des centromères et régions péri-centromériques des chromosomes. Il correspond à une chromatine très condensée, formant des kinétochores permettant l'ancrage des chromosomes aux microtubules durant la division cellulaire. L'hétérochromatine peut représenter une portion importante des génomes car elle correspond principalement aux zones pauvres en gènes, très peu actives transcriptionnellement mais riches en séquences répétées telles que les microsatellites et éléments transposables qui doivent être maintenus dans un état non transcrit (réprimés). Il est possible de retrouver également des zones d'hétérochromatine plus dispersées dans le génome qui correspondent aux régions intergéniques également riches en séquences répétées. La répartition de cet état chromatinien dépend donc essentiellement du contenu en TEs et de leur distribution le long des chromosomes.

L'euchromatine correspond à des zones riches en gènes et transcriptionnellement actives. Ce sont des zones de chromatine « ouvertes » et sont le lieu d'un remodelage fréquent avec une densité de nucléosomes moins importante (Strålfors et Ekwall 2011). Elle est associée à un nombre de marques épigénétiques beaucoup plus important permettant des combinaisons complexes et une multitude d'états fonctionnels différentiés. Ainsi, on distingue la région promotrice des gènes (promoteurs + TSS/5'UTR) caractérisée par la présence des marques H3K4me2/3 et les variants d'histones H3.3, H2A.Z. Selon l'abondance de ces variants et de la présence des marques H2Bup et H3K36me3, deux états chromatiniens sont distinguables au niveau des promoteurs (Vergara et Gutierrez 2017). Les régions géniques actives, avec les séquences codantes de la partie 5' UTR jusqu'à la partie 3' comprennent au niveau de la région 5' un enrichissement de H3K4me1/2, des niveaux faibles en H3K27me3 et un enrichissement en H3K4me2 dans la partie 3'. De plus, l'euchromatine transcriptionnellement active est souvent associée à des histones

acétylées avec notamment les marques H3K9ac et H3K56ac (Vergara et Gutierrez 2017). La marque H3K9ac est associée aux régions génomiques activement transcrites mais son rôle n'est pas précisément connu. Dans leur étude sur des cellules Hela humaines, Gates *et al*. 2017 ont démontré que le recrutement du complexe SEC (Super Elongation Complex) était conditionné par la présence de cette marque histone. De fait, elle pourrait être impliquée dans la phase d'élongation de la transcription de l'ADN chez l'homme. Par ailleurs, la revue de Hu *et al.* 2019, qui décrit tous les stress qui impactent les profils de H3K9ac chez diverses espèces de plantes, indique que sa présence est indispensable à l'adaptation des espèces via une transcription réactive efficace face aux aléas environnementaux.

L'hétérochromatine facultative correspond à un état chromatinien répressif transitoire de la transcription. Selon les stades et tissus de développement ou selon des changements de conditions environnementales, certaines régions intergéniques comprenant des motifs de régulation et certains gènes sont associées à la marque épigénétique H3K27me3, caractéristique de l'hétérochromatine facultative ou transitoire (Wiles et Selker 2017). Cet état chromatinien couvre 13,5% du génome d'Arabidopsis (Vergara et Gutierrez 2017). Il peut également cibler des TE présents dans les régions intergéniques des régions distales (Li *et al.* 2019). Cette marque est déposée de façon dynamique au cours du développement par un complexe protéique, appelé Polycomb, caractérisé en premier lieu chez la drosophile (Chittock *et al.* 2017). Ce Polycomb repressive complex 2 (PRC2) est conservé au sein des eucaryotes et est associée à l'identité cellulaire (Lavarone *et al.* 2019). L'accumulation de copies de gènes pour ce complexe de régulation chez les plantes démontre son importance dans les processus d'adaptation et d'évolution des végétaux (Buzas 2017).

III.1.3.2 **Distribution des marques épigénétiques**

Grâce à l'optimisation de la technique d'étude des marques histones (Chromatine ImmunoPrecipitation ou ChIP) qui consiste en l'immunoprécipitation des fragments d'ADN accrochés aux histones reconnues par un anticorps reconnaissant une marque histone spécifique, il est désormais possible de définir des états chromatiniens ayant des propriétés spécifiques. Ces états chromatiniens correspondent aux fréquences de combinaisons de marques aux loci génomiques présentant un pic de détection partagé par différentes marques. La première étude, pionnière dans ce domaine chez les plantes est celle de Roudier *et al*. 2011. Quatre états chromatiniens ont ainsi été

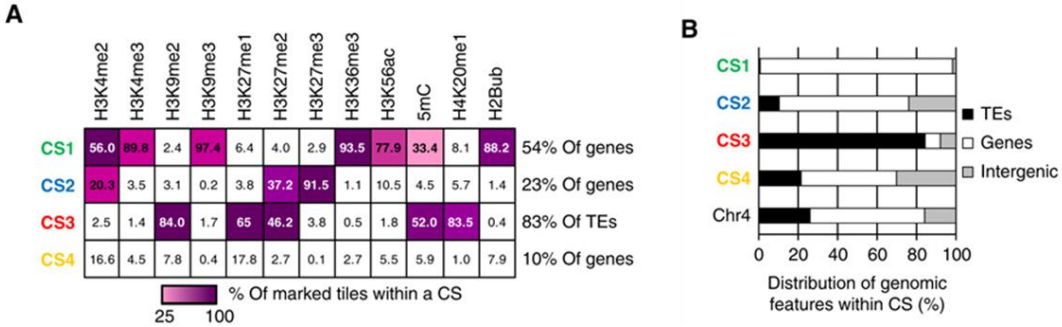


Figure 24. Présentation des 4 états chromatiniens identifiés chez Arabidopsis thaliana.

A)*Composition en marques épigénétiques des 4 états chromatiniens en étudiant 12 marques épigénétiques (CS : Chromatine State). La nuance de violet des cases indique les proportions majoritaires des marques définissant chacun des 4 CS.* ***B)*** *Proportion relative de chaque type de séquence nucléotidique pour chaque CS*

définis (CS1 à CS4) chez *Arabidopsis thaliana* en analysant 12 marques épigénétiques couvrant 90% du génome. L'état CS3 est constitué majoritairement de TE et correspondrait à de l'hétérochromatine constitutive, CS1 correspondrait à de l'euchromatine et à des gènes transcriptionnellement actifs et CS2 correspondrait plutôt à de l'hétérochromatine facultative (Figure 24).

Chez l'orge, 11 états chromatiniens ont été déterminés par l'analyse des distributions de 9 marques histones, chacun présentant des proportions plus ou moins importantes des différentes marques et associé à des types de séquences différents (Baker *et al.* 2015).

Les cartes génomiques de ces états chromatiniens constituent des outils intéressants pour corréler la dynamique de transcription des gènes et la régulation épigénétique à l'échelle du génome. Elles permettent d'explorer et caractériser la structuration fonctionnelle des chromosomes et de relier les données de marquage épigénétique de la chromatine à des données d'expression des gènes. Dans le cadre de l'étude des biais d'expression au sein des génomes des espèces polyploïdes l'établissement des paysages des différents états chromatiniens des sous-génomes constitue une nouvelle approche pour comprendre le fonctionnement et l'évolution d'un génome polyploïde (Sharma *et al.* 2018).

La distribution des marques histones peut également être étudiée à l'échelle du gène et des régions régulatrices adjacentes et définir avec précision les profils de marques au niveau des séquences codantes. La distribution le long des séquences codantes permet de corréler la présence de certaines marques à l'activité de transcription. La Figure 25 résume de façon schématique les profils des principales marques histones connues au niveau des gènes et leur rôle sur la transcription des gènes (Barth et Imhof 2010). Ainsi, la forme de la chromatine le long des chromosomes mais aussi au niveau des séquences codantes et régulatrices constitue une nouvelle source d'informations pour comprendre la régulation des gènes.

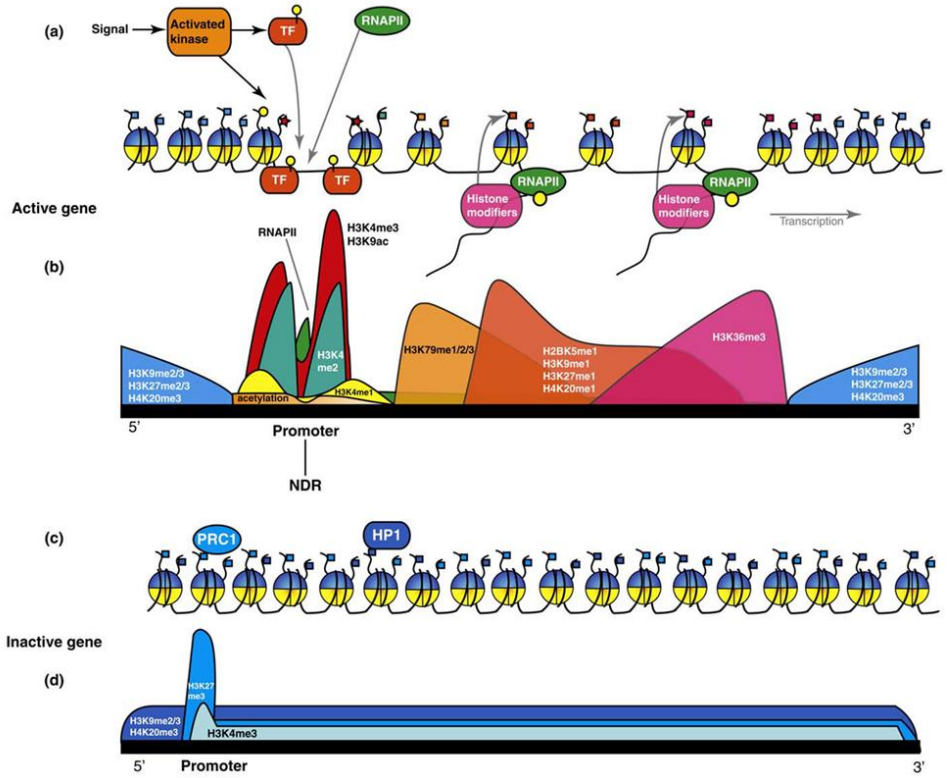


Figure 25. Distribution des marques histones le long d'un gène selon son état transcriptionnel (Barth et Imhof 2010).

III.1.4 **Epigénomique chez les espèces allopolyploïdes modèles**

La polyploïdisation implique tout autant des remaniements potentiels des transcriptomes que des remaniements touchant la distribution des territoires chromatiniens et des marques épigénétiques au niveau des séquences codantes. Que se passe-t-il lorsque deux génomes se retrouvent dans un même noyau ? Les patrons épigénétiques sont-ils remaniés et harmonisés ou y a-t-il une coexistence des états chromatiniens différents entre deux sous-génomes homéologues ? L'évolution des bais d'expression des gènes homéologues au cours de l'évolution peuvent-ils être, au moins en partie, expliqués par un remaniement des épigénomes ? Comme nous l'avons vu pour les réarrangements de séquences, les remaniements chromatiniens vont se produire selon deux scénarii, à des échelles de temps différentes : directement après allopolyploïdisation et tout au long de l'évolution du génome, notamment durant la phase de diploïdisation post-polyploïdisation.

Par exemple, le fractionnement biaisé de l'un des sous-génomes pourrait être lié à des différences en termes de régulation épignénétique chez les espèces diploïdes. Si l'un des deux sous-génomes présente un environnement chromatinien défavorable à l'expression de certaines séquences, relativement aux autres séquences homéologues, cela peut entraîner un relâchement de la pression de sélection et une perte progressive de ces séquences par pseudogénisation (Zhao *et al.* 2017, Whoodhouse *et al.* 2014, Schnable *et al.* 2011, Bottani *et al.* 2018). D'autres études semblent plus réservées quant à ce type de corrélation. Chez le maïs, par exemple, les profils globaux de la méthylation de l'ADN ou de la marque histone H3K27me3 ne sont pas drastiquement différents entre les sous-génomes 1 et 2 chez cette espèce paleo-polyploïde (Eichten *et al.* 2011, Makarevitch *et al.* 2013, West *et al.* 2014). De même, la sous-fonctionnalisation de gènes dupliqués pourrait s'expliquer par un différentiel de marquage épigénétique déjà présent chez les progéniteurs diploïdes ou acquis au cours de l'évolution de l'espèce polyploïde de par la redondance génétique (Song et Chen 2015).

Un nombre croissant d'études proposent la caractérisation de la réorganisation des patrons épigénétiques chez des espèces autopolyploïdes et allopolyploïdes : 68 études depuis 2013 sur plusieurs espèces de plantes ont déjà été réalisées (Tayalé et Parisod 2013). Si beaucoup d'études ont caractérisé le rôle de la méthylation de l'ADN, encore peu d'études sont disponibles pour comprendre le rôle et l'implication des marques histones et des états chromatiniens dans ces processus.

III.1.4.1 **Allopolyploïdisation et méthylation de l'ADN**

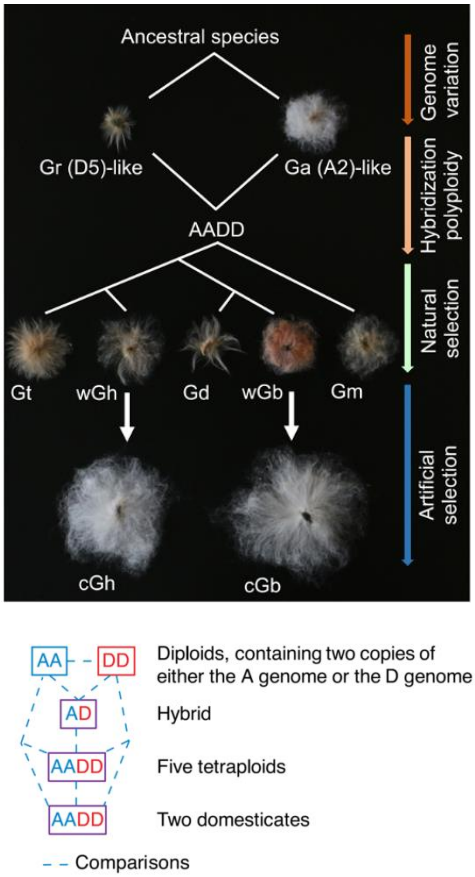
Dans de nombreuses études, ce sont le différentiel de la proportion en TEs entre les génomes progéniteurs et les différentes proportions des TE aux abords des gènes qui ont été identifié comme facteurs expliquant le différentiel d'expression entre les gènes homoeologues chez une espèce allopolyploïde voire le fractionnement biaisé de l'un des sous-génomes (Song *et al.* 2017 chez le coton, Edger *et al.* 2017 chez Mimulus, Zhao *et al.* 2017 chez le maïs et le soja, Cheng *et al.* 2016 chez *Brassica rapa*, Hollister *et al.* 2011 chez Arabidopsis). En effet, les TEs étant silencés d'un point de vue épigénétique par la méthylation

de l'ADN notamment et d'autres marques épigénétiques telles que la marque histone H3K9me2, la présence de ces derniers est supposée créer un environnement chromatinien défavorable à la transcription des gènes qui peut être différent entre deux sous-génomes homéologues (Vicient et Cascuberta 2017, Springer *et al.* 2016). En particulier, les méthylations de type CHH et la marque histone H3K9me2 sont supposées avoir un rôle dans le silencing des TE aux abords des gènes au niveau des frontières entre hétérochromatine et euchromatine (Gent *et al.* 2014, Li *et al.* 2015). Chez les espèces précédemment citées, ce phénomène est souvent corrélé à la dominance d'expression de l'un des sous-génomes (Woodhouse *et al.* 2014, Bottani *et al.* 2018, Bird *et al.* 2018).

Cependant, certains auteurs ont montré une relation inverse avec une modification de l'état épigénétique (méthylation CHH) des TEs au niveau de ces frontières à la suite de l'expression des gènes (Secco *et al.* 2015). Plusieurs études ont aussi démontré des interactions intergénomiques des sous-génomes chez des individus allopolyploïdes resynthétisées et identifié des changements stochastiques de méthylation à certains *loci*, notamment pour les espèces des gènes *Arabidopsis* et *Brassica* (Song et Chen 2015, Ding et Chen 2018). Enfin, le processus de répression des TEs près des gènes par une méthylation *de novo* via la voie RdDM est dissymétrique au sein d'un génome polyploïde. En effet, le génome maternel présente une méthylation prépondérante des TEs puisque les siRNA impliqués dans le silencing de type RdDM proviennent du cytoplasme du gamète femelle.

Deux études ont montré de façon très claire l'implication de la méthylation de l'ADN et notamment des TEs dans les biais d'expression des gènes :

Pour le genre *Mimulus*, chez l'hexaploïde *M. peregrinus*, la dominance d'expression du sous-génome Lb est reliée à une plus faible proportion de TEs aux abords des gènes de ce sous-génome, elle-même associée à une plus faible méthylation de l'ADN (Edger *et al.* 2017). Certaines études ont démontré une déméthylation des TEs pour les premières générations suivant la formation d'espèces allopolyploïdes, qui retrouveraient ensuite graduellement une méthylation correspondant à celles des progéniteurs (Parisod *et al.* 2009). Dans l'étude de Edger *et al.* 2017, les différences de méthyaltion des TEs seraient lié à une dissymétrie de ce processus. Chez les polyploïdes naturels (*M. peregrinus*) les densités de methylation CHH au niveau des TE pour le sous-génome G provenant de *M. gattatus* retrouvaient des niveaux proches de ceux du progéniteur alors que ceux de *M. luteus* restaient bas et au-dessous de ceux du progéniteur correspondant. Ce différentiel de méthylation impliquerait une transcription plus active des gènes de ce sous-génome moins riche en TE près des gènes.



Chez le coton, une étude similaire a été réalisée sur système biologique permettant d'identifier les réarrangements épigénétiques (méthylation de l'ADN) sur quatre phases évolutives du coton (Song *et al*. 2017). La Figure 26 ci-contre présente les espèces utilisées : deux ancêtres diploïdes (génomes AA et DD), un hybride synthétique (AD), quatre espèces allotétraploïdes sauvages (AADD) formées il y a 1-1,5 MYA et deux espèces allotétraploïdes domestiquées (AADD). Le principal résultat démontré est la conservation des profils de méthylation de l'ADN entre ces différentes espèces et donc leur héritabilité au cours de l'évolution. De plus, la diversité épi-allélique liée à la méthylation de l'ADN est plus fréquente que les variations nucléotidiques entre ces différentes accessions (les taux de substitution cytosines méthylées/non méthylées changeraint plus rapidement que les substitutions neutres (ks) au niveau des séquences), suggérant un rôle fort de la méthylation de l'ADN dans l'évolution des génomes. L'analyse des différences de

Figure 26. Système biologique utilisé pour étudier les conséquences épigénétiques de l'hybridation et de la polyploïdisation chez le coton.

En haut les photographies des inflorescences des 9 espèces apparentées au coton domestiqué. Espèces diploïdes progénitrices G. arboreum (Ga, génome A), G. raimondii (Gr, génome D), coton tetraploïde sauvage G. hirsutum (wGh), coton tetraploïde sauvage G. barbadense (wGb), G. tomentosum (Gt), G. darwinii (Gd), G. mustelinum (Gm), coton domestiqué G. hirsutum (cGh) et G. barbadense (cGb). Sur la droite, processus naturels et anthropiques donnant les espèces présentées.
En dessous, schéma des comparaisons réalisées lors de l'étude de la méthylattion de l'ADN, du transcriptome, des divergences de séquence entre autres. Photographies issues de Song et al. 2017 ; Schéma issu de Jackson 2017

contenu en éléments transposables entre les sous-génomes des espèces polyploïdes a révélé que le sous-génome A est plus grand car il présente un pourcentage de TE plus important et de fait une proportion de cytosines méthylées deux fois plus importante comparé au sous-génome D. En revanche, ce dernier présente des densités de mCG et mCHG dans le corps des gènes plus élevées. L'ensemble de ces différences expliquerait une partie des biais d'expression des gènes homéologues observés au sein des génomes polyploïdes, à savoir une expression du sous-génome A plus importante que celle du sous-génome D. L'hypothèse synthétique proposée par les auteurs pour expliquer le différentiel d'expression global entre les deux sous-génomes homéologues serait que le sous-génome A présenterait une compartimentation plus importante des TE au niveau des régions péri-centromériques et une remobilisation moins importante de ceux-ci au niveau de l'espace génique comparativement au sous-génome D. Ceci limiterait la présence de TE au niveau des gènes, dont la séquence serait couverte de méthylations inhibitrices de l'expression, contrairement à ce qui a été observé pour le sous-génome D.

Ainsi, ces différents travaux montrent le lien étroit entre les caractéristiques des portions d'éléments transposables au sein des génomes polyploïdes, les variations de méthylation de l'ADN associées et l'expression différentielle entre deux sous-génomes ou entre *loci* homoeologues.

III.1.4.2 **Allopolyploïdisation et marques histones**

Les analyses de l'implication des marques histones dans les différences d'expression des gènes homéologues sont beaucoup plus récentes, moins nombreuses et concernent surtout des loci spécifiques. De plus, du fait de la difficulté de la produciton de données de ChIP-seq, les études ne concernent pour l'instant très souvent qu'une ou deux marques à la fois et non une combinaison de marques. A l'heure actuelle, peu de travaux sur les marques histones ont été reliés à la dominance d'expression de l'un des sous-génomes ou au fractionnement des génomes.

Chez le coton, des différences de densité de la marque H3K4me3 ont été identifiées entre *loci* hoémologues sur des cellules de racines pour l'espèce allotétraploïde *Gossypium hirsutum* (Zheng *et al*. 2016). Même si les niveaux de densité de cette marque sur l'entièreté des chromosomes des deux sous-génomes A et D étaient sensiblement les mêmes, certaines régions chromosomiques présentaient des densités très différentes (fold change de 2). Ces différences ont été corrélées à des différences d'expression entre les gènes homéologues des deux sous-génomes aux loci considérés. Certaines régions riches en gènes du sous-génome D, exhibant un niveau d'expression plus faible que les régions homéologues du sous-génome A, présentaient une densité beaucoup plus faible de H3K4me3.Il a également été également été mis en évidence que deux des 26 chromosomes présentaient des patrons particuliers d'expression des gènes qu'ils ont attribués aux translocations présentes sur ceux-ci, ayant probablement interverti ou modifié les territoires chromatiniens des deux chromosomes.

Certaines études traitent de la question de la sous-fonctionnalisation par variations épigénétiques mais une synthèse claire reste complexe à établir puisque les connaissances sont en cours d'acquisition.

Une étude réalisée chez Arabidopsis, comprenant les progéniteurs autopolyploïdes *A. thaliana* and *A. arenosa* dont l'hybridation interspécifique a donné l'allotetraploide *A. suecica,* avait identifié le rôle de certaines modifications histones dans les différences d'expression du locus FLC intervenant dans la transition florale (Wang *et al.* 2006). L'expression plus élevée du gène FLC chez l'allotetraploïde, entraînant une floraison plus précoce que celle de ses deux parents (FLC réprime la transition florale) a été corrélée à des densités de H3K9ac et H3K4me2 plus importantes au niveau du promoteur du gène.

Le rôle potentiel de la marque H3K27me3 dans l'évolution des profils d'expression des gènes à la suite d'une hybridation, intéressants pour comprendre les potentiels changements lors d'une allopolyploïdisation, a été étudié dans un système impliquant des espèces du genre *Arabidopsis*. Chez l'hybride F1 issu d'un croisement des espèces *A. lyrata* et *A. thaliana* il a été montré une conservation (au lieu d'un remaniement) du marquage H3K27me3 entre les espèces parentales et l'hybride dans les premières générations (Zhu *et al.* 2017). En revanche, les gènes s'exprimant davantage dans le génome

hybride, en comparaison du génome parental (up-regulated), présentent une perte du marquage H3K27me3. Les auteurs ont aussi mis en évidence que les biais d'expression (dominance globale d'expression des gènes d'*A.Lyrata* sur ceux d'*A.thaliana*, avec une grande majorité de gènes d'*A.thaliana* réprimés) étaient davantage corrélés à un différentiel de compaction de la chromatine entre les deux sous-génomes au sein de l'hybride plutôt qu'aux différents patrons de méthylation de l'ADN.

Toujours chez *Arabidopsis thaliana*, une étude sur 3169 paires de gènes paralogues marqués H3K27me3 avait démontré que deux paralogues marqués H3K27me3 présentaient peu de divergence d'expression et de divergence de séquence au niveau des régions cis-régulatrices mais qu'ils avaient des taux les plus élevés de divergence de séquences codantes (corps des gènes) comparé aux paralogues non marqués H3K27me3 (Berke *et al.* 2012). En revanche, les paires de paralogues dont seulement l'un des deux gènes était marqué H3K27me3 présentaient des divergences d'expression et de séquences régulatrices très marquées. Ainsi, selon cette étude, la présence de cette marque contraindrait l'évolution des gènes dupliqués. Cette étude montre qu'il est crucial d'étudier en parallèle les divergences de séquences et les divergences de marquage épigénétique pour ne pas conclure à l'implication unique des variations épigénétiques quant à l'évolution des variations d'expression des gènes dupliqués.

Dans une étude chez *Brassica rapa* sur deux lignées parentales donnant l'hybride F1 et sur deux tissus différents (cotyledons et feuilles 14 jours après semis) il a été montré que les gènes paralogues au sein du génome hybride qui ont des biais d'expression entre les deux tissus présentaient un enrichissement de la marque H3K27me3 (Akter *et al.* 2019).

Une autre façon de connaître la participation potentielle de la régulation épigénétique dans l'évolution des fonctions des gènes est d'étudier le partage des marques épigénétiques entre espèces proches.

Une comparaison des localisations génomiques de H3K27me3 entre quatre espèces de drosophiles et avec celles du ver *Caenorhabditis elegans*, en distinguant les gènes dupliqués en tandem des gènes dupliqués et dispersés dans le génome, a permis de mettre en évidence une conservation globale du marquage entre espèces sauf pour les gènes dupliqués dispersés dans le génome, comparé aux gènes issus d'autres types de duplications (Arthur *et al.* 2014). Ils ont corrélé l'implication du marquage épigénétique H3K27me3 dans l'évolution des patrons d'expression des gènes dupliqués. Ils ont aussi relié ces divergences d'expression avec des divergences de séquence, les gènes dupliqués exprimés différemment étant marqués H3K27me3 présentaient des ratios dN/Ds (équivalent du ratio ka/ks) plus élevés (évolution de la fonction de la séquence).

Une étude similaire réalisée sur trois espèces d'Arabettes ayant divergées il y a 24 millions d'années et 6 millions d'années (*Arabidopsis alpina* vs *Arabidopsis lyrata* et *Arabidopsis thaliana* et les deux dernières espèces entre elles respectivement) a défini deux types de gènes orthologues marqués H3K27me3 : les gènes « plastiques » et les gènes « contraints » (Chica *et al.* 2017). Les derniers présentaient une conservation du marquage entre *Arabidopsis alpina* et *Arabidopsis lyrata* et/ou *Arabidopsis thaliana* alors

que les gènes orthologues « plastiques » présentaient le marquage seulement chez l'une des deux espèces *Arabidopsis lyrata* et *Arabidopsis thaliana* ou uniquement chez *A. alpina*. Les auteurs ont montré que les gènes H3K27me3 « contraints » présentaient une divergence de séquence moindre comparée aux gènes dits « plastiques ». De plus, d'après leurs analyses de données RNA-seq sur les trois espèces, les gènes H3K27me3 contraints présentaient une spécificité d'expression tissulaire plus importante. Ainsi, la conservation du marquage H3K27me3 contraindrait l'évolution des séquences promotrices et donc l'évolution fonctionnelle des gènes portant la marque H3K27me3.

Ces différentes études montrent l'implication potentielle des marques épigénétiques de types modifications postraducitonnelles des histones pouvant expliquer les divergences d'expression des gènes homéologues au sein d'un génome polyploïde, en particulier pour la marque H3K27me3. Cependant, des différences d'évolution de séquences sont aussi retrouvées pour les gènes marqués. Il est donc encore difficile de clairement définir les épimutations comme étant la causalité unique de l'évolution de la régulation des gènes puisqu'il faut pour cela systématiquement prouver que ces changements ne sont pas liés à des polymorphismes de séquence.

Même si les liens entre processus épigénétiques et évolution des génomes polyploïdes ne sont pas encore très clairs, ils semblent jouer un rôle dans ce processus. Leur compréhension permettra de mieux caractériser le phénomène de polyploïdisation qui semble avoir été sélectionné au cours de l'évolution des espèces végétales puisqu'il se répéter de façon plus ou moins aléatoire, notamment chez les espèces domestiquées par l'homme.

III.1.5 **Régulation épigénomique chez le blé tendre**

Le séquençage du génome du blé tendre étant relativement récent, les analyses à l'échelle du génome entier sur les aspects épigénétiques sont encore peu nombreuses. Les premières études ont concerné la méthylation de l'ADN avec notamment l'étude de Gardiner *et al*. 2015. Les auteurs ont analysé les profils de méthylation de 81Mb de régions géniques chez le blé allohexaploïde dans deux contextes thermiques de condition de cultures : 12°C et 27°C. L'objectif de l'article était de caractériser les régions géniques susceptibles de présenter une méthylation différentielle entre loci homéologues selon les variations de température. Les auteurs n'avaient identifié que 2,4% de méthylation différentielle entre les trois sous-génomes et les deux températures, avec 45% des séquences présentant les mêmes niveaux de méthylation entre A, B et D. Les auteurs ont par ailleurs corrélé les niveaux d'expression plus faibles de certains gènes à des patrons de méthylation différenciés entre les promoteurs. Enfin, ils ont démontré en utilisant des séquences du progéniteur *Ae. Tauschii* que les patrons de méthylation étaient globalement conservés entre l'espèce polyploïde et l'espèce diploïde (seulement 4,8% des régions présentaient des variations de méthylation détectées sur 14Mb comparées).

La Figure 27 résume l'ensemble des résultats concernant les profils de méthylation pour les deux conditions testées et les trois sous-génomes.

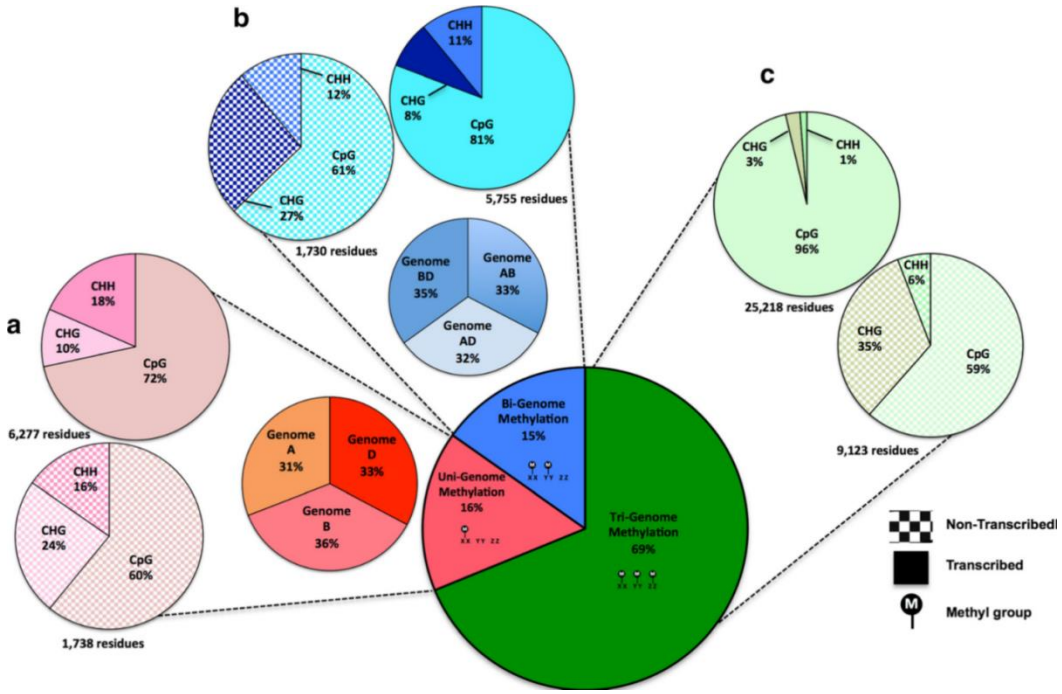


Figure 27. Proportion de loci homéologues méthylés selon les trois contextes CG, CHG, CHH chez le blé tendre dans deux conditions de température,12°C et 27°C.

Cercle du milieu : proportions globales des méthylations sur un, deux ou les trois loci homéologues.
***a)** Méthylation unique d'un des sous-génomes, avec le détail des types de méthylation CG, CHG, CHH pour les régions transcrites, non-transcrites/promoteurs*
***b)** Méthylation partagée par deux des sous génomes avec le détail des types de méthylation CG, CHG, CHH pour les régions transcrites, non-transcrites/promoteurs*
***c)** Méthylation partagée par les trois sous-génomes, conservation des profils de méthylation, avec le détail des types de méthylation CG, CHG, CHH pour les régions transcrites, non-transcrites/promoteurs. Gardiner et al. 2015*

A l'échelle d'un locus en particulier, Hu et ses collaborateurs (2013) ont étudié l'expression du gène, *TaEXPA1* de la famille des Expansines, connu pour contrôler la taille des cellules dans les tissus feuilles et racines. Ils ont analysé l'expression de ce gène chez le blé hexaploïdie et les trois espèces diploïdes pour mieux comprendre le rôle de la polyploïdisation dans l'évolution de l'expression de chacun des homéologues. Ils ont mis en évidence que les trois copies sont éteintes dans les racines des plantules et seules les copies A et D sont exprimées dans les feuilles. Grâce à des analyses ChIP-seq, ils ont montré que dans les racines, les densités de la marque H3K9me2 étaient élevées et celles des marques H3K4me3 et H3K9ac étaient faibles, sur les trois copies du gène considéré. La réactivation de l'expression des copies A et D au niveau des feuilles a été corrélée à un inversement des densités des marques activatrices et inhibitrices de la transcription par rapport au tissu racine. En analysant la méthylation des cytosines sur le gène et son promoteur chez les génotypes tétraploïdes et hexaploïdes naturels et synthétiques, les auteurs ont également démontré une méthylation plus importante des cytosines pour la copie B au sein du polyploïde naturel pouvant induire sa répression.

De même, l'étude de Zhang *et al*. 2017 sur le gène TaGS2 (GS= glutamine synthase cruciale pour l'assimilation de l'ammonium) met en évidence les changements chromatiniens reliés à l'expression différentielle des homéologues de ce gène chez le blé. D'un point de vue de l'expression, le gène présente des niveaux d'expression (q-PCR) plus élevés chez l'espèce diploïde donnant le sous-génome B comparé aux deux autres progéniteurs. Chez l'hexaploïde, la contribution relative de chaque homéologue pour l'expression totale du locus au sein du tissu feuille des plantules de 10 jours est de 75% pour la copie B, 23% pour la copie A et 2% pour la copie D. Les auteurs ont vérifié si ces différences ne pouvaient être reliées à des divergences de séquence. Ils n'ont trouvé aucune différence au niveau des séquences promotrices et l'analyse des séquences protéiques a révélé 99.84% d'identité de séquence entre les trois copies homéologues, avec deux acides aminés de divergence entre A, D et B. En revanche, les 1000pb en amont du gène comprennent des variations de séquences beaucoup plus nombreuses (82% de similarité de séquence calculés) mais les auteurs n'ont pas exploré les potentiels changements dans les motifs de régulation *cis*. Les auteurs ont réalisé un ChIP-seq de H3K4me3 et ont démontré un enrichissement de cette marque au niveau des premiers codons transcrits (TSS : Transcription Starting Site) pour la copie TaGS2-B comparativement aux copies A et D ce qui pourrait expliquer la dominance d'expression de cette copie. Cette étude est originale pour s'être également intéressée à la forme de la chromatine en elle-même en étudiant l'accessibilité de la chromatine. Pour cela, les auteurs ont observé les degrés de dégradation des promoteurs et séquences géniques des trois copies par des enzymes coupant l'ADN (DNAase1 et miccrococal nucléase). Ils ont mis en évidence une dégradation moins rapide des régions codantes des copies A et D par rapport à celles de la copie B. Cette étude confirme donc l'intérêt des études épigénétiques pour comprendre la sous-fonctionnalisation des copies par des mécanismes épigénétiques au sein d'un génome polyploïde.

Lors de la publication de la séquence de référence du génome en 2018, l'équipe de Moussa Benhamed (IPS2, Paris Saclay) a produit les premières données de ChIP-seq des marques histones H3K4me4, H3K27me3, H3K36me3, H3K9ac sur un tissu (feuilles, stade trois feuilles, environ dix jours après semis) (IWGSC 2018). L'analyse de la distribution de ces marques sur les chromosomes a révélé un partitionnement spécifique avec un enrichissement des marques associées à une transcription active des gènes (H3K36me3, H3K9ac, H3K4me4) dans les régions proximales R2a et R2b et un enrichissement de la marque H3K27me3 au niveau des régions distales. Ces résultats suggèrent un partitionnement des états chromatiniens le long des chromosomes qui serait corrélé au partitionnement des caractéristiques d'expression des gènes (gènes à amplitude et niveau d'expression plus élevés dans les régions proximales, gènes à expression et fonction tissu/condition spécifique dans les régions distales).

La récente étude publiée par Li *et al.* 2019 complète ces premiers travaux en publiant des données issues de sept ChIP-seq pour les marques histones H3K4me3, H3K27me3, H3K9ac et H3K27ac, H3K4me1, H3K9me2, H3K36me3 sur le même tissu que l'étude précédente c'est-à-dire les feuilles (stade trois feuilles). Ils ont également étudié l'ouverture de la chromatine ainsi que le méthylome. Ils ont observé le même partitionnement des marques histones le long des chromosomes observé dans IWGSC 2018 ainsi

qu'un enrichissement de la marque H3K9me2 et de la méthylation de l'ADN au niveau des centromères et régions proches, caractéristiques des régions riches en TEs. Les auteurs ont défini cinq groupes de gènes présentant différentes marques histones et les ont associés aux patrons d'expression des gènes (Figure 28).

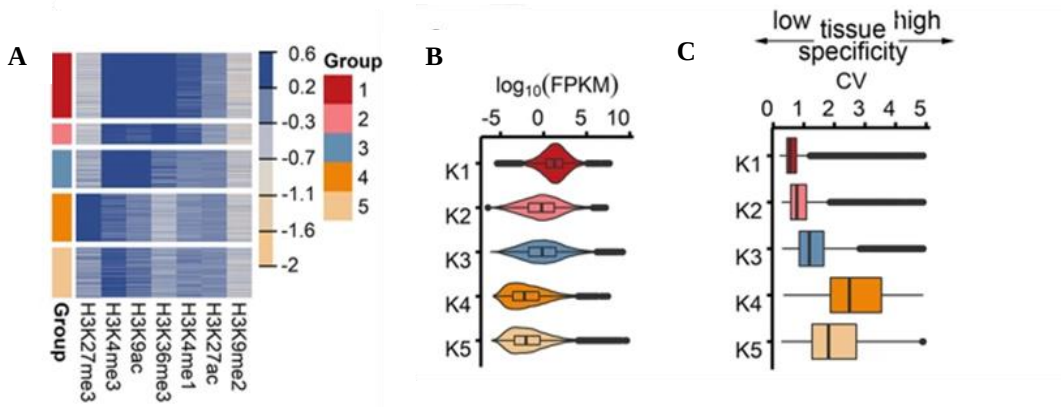


Figure 28. Groupes de gènes définis selon les combinaisons de marques histones chez le blé tendre.
***A**) Cinq groupes de gènes ont été définis selon les marques histones présentes sur les promoteurs et le corps des gènes.*
***B**) Distribution des densités d'expression (log10 de la moyenne de l'expression des gènes (en FPKM), violin plots)*
***C**) Spécificité d'expression des gènes (calcul du coefficient de la variance d'expression (CV) dans sept tissus (données RNA-seq Ramirez-Gonzales 2018). Stade 3 feuilles. 2 réplicas biologiques. Li et al. 2019.*

Ils ont constaté que les gènes associés au groupe 1 (H3K4me3, H3K9ac, and H3K36me3) présentaient des niveaux d'expression plus élevés que ceux du groupe 4 (H3K27me3) qui eux arboraient une expression tissu-spécifique. Ce résultat reste concordant avec le rôle du complexe polycomb dans l'établissement de patrons de régulation des gènes spécifiques à l'identité cellulaire.

Tout comme les études de l'expression différentielle des homéologues chez le blé tendre présentées dans les chapitres 1 et 2 de cette thèse, les auteurs ont souhaité identifier les divergences de marques épigénétiques entre les trois génomes homéologues pour les gènes présentant un ratio 1 :1 :1. Ils ont surtout observé de plus grandes variations de densités des marques activatrices H3K4me3, H3K9ac, et H3K36me3 au sein des groupes de gènes homéologues. Ils ont corrélé significativement ces biais de densités de lectures aux divergences d'expression des gènes de ces groupes.

Enfin, ils ont défini 15 états chromatiniens par l'analyse des combinaisons de marques à l'échelle du génome qu'ils ont associés aux patrons d'expression et à la divergence de séquence entre régions homéologues. L'état chromatinien 4-5, enrichi en marques H3K9ac et H3K27ac représente 1,5% des séquences du génome correspondant aux gènes exprimés dans ce tissu. Ils ont surtout démontré une conservation du marquage des différents états chromatiniens entre les trois sous-génomes témoignant, tout comme pour les données d'expression, d'une absence de remaniement drastique des régulations épigénétiques entre les trois sous-génomes. Ainsi, les études épigénomique à l'échelle du génome entier manquent encore pour pouvoir corréler de façon robuste et claire les profils d'expression des gènes homéologues et les patrons épigénétiques potentiellement associés. En particulier, aucune étude épigénomique de ChIP-seq n'a été réalisé sur un ensemble de tissus afin d'obtenir un atlas de régulation épigénétique au cours du développement.

IV. Objectifs de la thèse

Les espèces végétales présentent des caractéristiques particulières du fonctionnement et de l'évolution de leur génome liées aux nombreux évènements de polyploïdisation réccurents dans leurs lignées évolutives. Selon l'âge de formation du polyploïde (paleo/mezo/neo allopolyploide) et le degré de divergence (séquence, niche écologique) des espèces progénitrices, on observera des patrons plus ou moins marqués (i) de conservation et de rétention de gènes selon l'**hypothèse de l'équilibre de dose**, (ii) un **fractionnement biaisé ou stochastique** entre les sous-génomes lié ou non à des **biais d'expression** (« use it or lose it »), (iii) de **neo-sous-fonctionnalisations** et **rétention de gènes dupliqués**. La présence et **la proportion de TE aux abords des gènes** entre sous-génomes ont été identifiées comme pouvant expliquer les biais d'expression entre gènes et/ou sous-génomes homéologues et le fractionnement biaisé de l'un d'entre-eux. De même, l'évolution de la redondance génique vers de la diversité fonctionnelle pourrait être expliquée par des différences originelles entre espèces diploïdes progénétrices ou l'acquisition de patrons de régulation épigénétique, avec en particulier les **marques histones** qui définissent des **environnements chromatiniens spécifiques**.

Le **blé tendre** est une espèce **neo-allopolyploïde** formée par deux hybridations successives il y a 800 000 et 10 000 ans, la première donnant le génome AABB tétraploïde et la dernière le génome hexpaloïde AABBDD. Il présente une redondance génique élevée liée aux deux évènements de polyploïdisation mais aussi liée à une **présence importante de gènes en copies multiples** dans le génome (notamment pour le sous-génome B) en comparaison avec d'autres espèces de céréales.

Dans ce contexte, les questions de travail pour la caractérisation du fonctionnement et de l'évolution de ce génome sont :

- Les sous génomes A et B qui ont coévolué plus longtemps ensemble présentent-ils davantage de patrons de sous-fonctionnalisation et/ou une perte de séquences plus marquée que le sous-génome D ?
- Le sous-génome D est-il encore autonome en termes d'expression ? Présente-t-il encore des niveaux et profils d'expression de l'espèce diploïde progénitrice qui pourraient impacter la stoéchiométrie des protéines des deux autres sous-génomes ?
- La présence de gènes en copies multiples implique-t-elle une forte proportion de biais d'expression entre gènes homéologues liée à des sous-fonctionnalisation anciennes ?
- Le différentiel de proportion en TE des trois sous-génomes aux abords des gènes peut-il expliquer les biais d'expressions observés ?
- Observe-t-on des sous-fonctionnalisations et retentions de paralogues stochastiques ou préférentielles selon les sous-génomes ?
- En comparaison avec d'autres espèces polyploïdes, les patrons d'expression des gènes homéologues n'ont pas pu évoluer sur de longues périodes de temps chez le blé tendre ; les patrons de régulation épigénétique, qui représentent la part dynamique et réactive de l'ajustement du

fonctionnement d'un génome, permettent-ils de déceler les premiers évènements de divergence de régulation des gènes liés à la polyploïdisation ?

L'obtention récente d'une séquence de référence du génome complet pour le blé hexaploïde ouvre la voie à des analyses « whole genome » sur l'ensemble des gènes annotés. L'obtention des positions physiques précises des gènes et la détermination de leurs relations d'homéologie offre la possibilité d'explorer plus facilement les relations et les différences entre sous-génomes. Cela permet aisni d'identifier des tendances de biais d'expression, de régulation et de réarrangement génomique post-polyploïdisation.

Dans ce contexte scientifique et selon les hypothèses de travail, l'**objectif général de ce projet** était de caractériser le comportement des gènes homéologues en termes d'expression en fonction du nombre de copies de gènes dans le génome du blé tendre. Nous souhaitions également relier les tendances observées aux caractéristiques structurales et épigénétiques des chromosomes afin de proposer de nouvelles hypothèses pour comprendre l'évolution de la redondance génétique au sein de ce génome.

Trois axes de travail ont structuré ce travail doctoral et ce manuscrit de thèse :

- **Le premier** axe visait à étudier les biais d'expression et les patrons épigénétiques de groupes de gènes homéologues présentant 3 copies de gènes (1 :1 :1 gènes en triades). Le travail sur ces groupes est présenté dans le chapitre II. Il a été réalisé dans le cadre d'une collaboration avec l'équipe de Christobal Uauy (John Innes Center, Norwich, Angleterre) qui a compilé 850 échantillons de données RNA-seq pour réaliser un atlas de données d'expression sur plusieurs tissus et conditions d'expérimentations. Nous avons également collaboré avec l'équipe de Moussa Benhamed de l'IPS2 (Institute of Plant Science, Paris-Saclay) pour l'analyse des données de ChIP-seq obtenues pour les marques histones H3K36me3, H3K9ac, H3K27me3, H3K4me3 et notamment avec Lorenzo Concia, bio-informaticien de l'équipe.

- **Le second** axe a été de compléter ce premier travail en réalisant une étude sur des groupes d'homéologues présentant un nombre de copies s'écartant du ratio théorique 1 :1 :1 : les groupes avec 2 (gènes en dyades 0:1:1 - 1:0:1 - 1:1:0) et 4 copies (2:1:1 - 1:2:1 - 1:1:2). Nous avons sélectionné ces deux catégories de gènes pour explorer les deux phénomènes opposés de perte de séquence et de rétention de gènes dupliqués. Nous avons utilisé les mêmes données que pour l'article précédent.

Le troisième axe de recherche, plus méthodologique, visait à produire des données ChIP-seq pour la marque histone H3K27me3 chez *Triticum aestivum* cv Chinese spring. En effet au début du doctorat, très peu de données sur les marques épigénétiques de type modification postraductionnelles des histones étaient disponibles à l'échelle « whole genome » chez cette espèce. L'objectif initial était de produire un ensemble de données sur plusieurs tissus prélevés au cours d'une cinétique de développement, avec en parallèle la production de données RNA-seq afin d'obtenir un double atlas ChIP-seq/RNA-seq. Grâce à ces données, nous avions prévu d'étudier le marquage différentiel de la marque H3K27me3 des gènes homéologues au cours du développement. Je présenterai certains verrous qui ne m'ont pas permis d'atteindre cet objectif et les réflexions proposées pour les futures expérimentations.

CHAPITRE II

Paysage transcriptionnel chez le blé tendre : analyse des groupes homéologues en triades (1 :1 :1)

Article I : R. H. Ramírez-González, P. Borrill, D. Lang, S. A. Harrington, J. Brinton, L. Venturini, M. Davey, J. Jacobs, F. van Ex, A. Pasha, Y. Khedikar, S. J. Robinson, A. T. Cory, T. Florio, L. Concia, C. Juery, H. Schoonbeek, B. Steuernagel, D. Xiang, C. J. Ridout, B. Chalhoub, K. F. X. Mayer, M. Benhamed, D. Latrasse, A. Bendahmane, International Wheat Genome Sequencing Consortium, B. B. H. Wulff, R. Appels, V. Tiwari, R. Datla, F. Choulet, C. J. Pozniak, N. J. Provart, A. G. Sharpe, E. Paux, M. Spannagl, A. Bräutigam, C. Uauy The wheat transcriptional landscape. (2018), *Science* 17, Aug 2018 : Vol. 361, Issue 6403, DOI:10.1126/science.aar6089

Contexte

Les analyses d'expression des gènes chez le blé tendre ce sont faite en l'abscence d'une séquence génomique de référence jusqu'au 2018. En 2014, une première ébauche avait été produite avec un séquençage Illumina des chromosomes triés (« *Chromosome survey sequences* », IWGSC 2014) et avait permis des analyses transcriptomiques à l'échelle du génome entier (Pfeifer *et al.* 2014) mais la qualité de l'assemblage restait encore à améliorer. C'est au début de mon doctorat que la publication de la séquence de référence du génome complet du blé tendre reconnue par al communauté scietifique était en préparation au sein de l'IWGSC. En parallèle de ce travail, l'équipe de Christobal Uauy du John Innes center ainsi que des collaborateurs ont travaillé sur la compilation et la publication d'un atlas de données RNA-seq provenant de 850 échantillons prélevés à différents stades de développement pour les accessions de blé tendre Chinese spring et Azhurnaya. Ces données de RNA-seq ont permis de proposer la première étude complète de l'expression des gènes du blé tendre à l'échelle du génome entier. L'objectif de cet article était d'étudier le degré d'intégration ou d'autonomie des trois sous-génomes en termes de régulation de l'expression des gènes et le comportement relatif des gènes homéologues. L'intégration de données structurales et de données épigénétiques à ces analyses de transcriptomique permettraient de déterminer les mécanismes impliqués dans les profils et tendandes observées sur l'expression des gènes.

Stratégie

En se concentrant uniquement sur les gènes en triplets : 1 :1 :1, soit une copie par sous génome, 123 échantillons représentant 15 tissus et stades de développement ont été utilisés pour étudier les biais d'expression de ces gènes. Ils ont sélectionné les gènes présentant une moyenne d'expression sur les 3 réplicas biologiques pour chacun des tissus supérieure à 0,5TPM pour au moins un des 15 tissus/stades et ont sélectionné les groupes de triplets d'homéologues présentant une somme de leur expression supérieure à 0,5TPM. 119 échantillons pour les grains, 245 échantillons pour les feuilles, 45 échantillons pour les racines et 128 échantillons pour les épis ont également été utilisés pour construire des réseaux de co-expression des gènes grâce à une analyse WGCNA (Weighted correlation network analysis). Les données d'expression des gènes ont été corrélées aux données de structure des chromosomes (partitionnement structural, présence de TE dans les promoteurs) aux données d'épigénétique de méthylation de l'ADN et de marques histones (H3K4me3, H3K27me3, H3K36me3, H3K9ac).

Conclusions

Les principales conclusions de cet article sont que 70% des triades qui représentent ~50% des gènes du génome présentent une expression équilibrée entre les trois sous génomes. Les autres 30% présentent un biais d'expression majoritairement par sous-expression d'une (ou de deux) copies relativement aux autres gènes du groupe de gènes homéologues. Les auteurs ont mis en évidence que ces triades étaient plus fréquemment trouvées dans les régions distales des chromosomes, plus riches en événementde recombinaison et en marque histone H3K27me3. Ils ont également corrélé les forts taux d'expression des triades « balanced » (équilibrées) en termes de biais d'expression à des densités de marques histones H3K9ac, H3K4me3 et H3K36me3. Les triades équilibrées sont préférentiellement retrouvées dans les régions proximales des chromosomes, peu recombinogènes. De même une expression globale de l'ensemble des gènes du sous-génome D légèrement plus élevée a été corrélée à un marquage H3K27me3 légèrement moins important. Enfin, ils ont mis en évidence que les variations de biais d'expression entre les tissus, au cours du développement pour les 10% des triades les plus dynamiques pourraient être liées à différences de séquence notamment en termes de présence de TE aux abords des promoteurs.

Enfin, des réseaux de co-expression de gènes ont été créés : 51 à 78 modules d'expression pour les 4 réseaux correspondant aux gènes exprimés dans : les racines, les épis, les feuilles et les grains. Les données ont été utilisées pour démontrer leur intérêt dans la recherche de potentiels gènes candidats à travers l'analyse des relations entre gènes intervenant dans les différents processus résumés par les 51 à 78 modules comprenant 42 à 88% de tous les gènes exprimés dans chaque tissu.

Les données de réseau de co-expression de gène ont été utilisées pour démontrer leur intérêt dans la recherche de potentiels gènes candidats à travers l'analyse des relations entre gènes intervenant dans les différents processus résumés par les 51 à 78 modules comprenant 42 à 88% de tous les gènes exprimés dans chaque tissu.

Implication personnelle

Lorsque j'ai commencé ma thèse, les données de ChIP-seq sur les marques histones chez le blé venaient tout juste d'être obtenues par l'équipe de Moussa Benhamed à l'IPS2. J'ai ainsi pu participer au traitement bio-informatique des données pour me former pour mes propres expériences en testant par exemple des paramètres d'alignement et de peak-calling (détection des pics d'enrichiseement de lectures de séquençage pour les différentes marques histones). Par ailleurs, grâce aux données du partitionnement des chromosomes obtenues dans mon équipe d'accueil, j'ai pu participer aux analyses permettant de produire les données de distribution des marques histones le long des chromosomes.

RESEARCH ARTICLE

WHEAT GENOME

The transcriptional landscape of polyploid wheat

R. H. Ramírez-González[1]*, P. Borrill[1]*†, D. Lang[2], S. A. Harrington[1], J. Brinton[1], L. Venturini[3], M. Davey[4], J. Jacobs[4], F. van Ex[4], A. Pasha[5], Y. Khedikar[6], S. J. Robinson[6], A. T. Cory[7], T. Florio[1], L. Concia[8], C. Juery[9], H. Schoonbeek[1], B. Steuernagel[1], D. Xiang[10], C. J. Ridout[1], B. Chalhoub[11], K. F. X. Mayer[2,12], M. Benhamed[8], D. Latrasse[8], A. Bendahmane[8], International Wheat Genome Sequencing Consortium[13]‡, B. B. H. Wulff[1], R. Appels[14], V. Tiwari[15], R. Datla[10], F. Choulet[9], C. J. Pozniak[7], N. J. Provart[5], A. G. Sharpe[16], E. Paux[9], M. Spannagl[2], A. Bräutigam[17]§, C. Uauy[1]†

The coordinated expression of highly related homoeologous genes in polyploid species underlies the phenotypes of many of the world's major crops. Here we combine extensive gene expression datasets to produce a comprehensive, genome-wide analysis of homoeolog expression patterns in hexaploid bread wheat. Bias in homoeolog expression varies between tissues, with ~30% of wheat homoeologs showing nonbalanced expression. We found expression asymmetries along wheat chromosomes, with homoeologs showing the largest inter-tissue, inter-cultivar, and coding sequence variation, most often located in high-recombination distal ends of chromosomes. These transcriptionally dynamic genes potentially represent the first steps toward neo- or subfunctionalization of wheat homoeologs. Coexpression networks reveal extensive coordination of homoeologs throughout development and, alongside a detailed expression atlas, provide a framework to target candidate genes underpinning agronomic traits in wheat.

Polyploidy arises from whole-genome dup-lication or interspecific hybridization and is ubiquitous in eukaryotic plant and fun-gal lineages. Polyploidy has been proposed to confer adaptive plasticity, thereby shaping the evolution of plants, fungi, and, to a lesser degree, animals (1, 2). This plasticity has facilitated the domestication and adaptation of several major crop species (3), including hexaploid bread wheat (Triticum aestivum; AABBDD subgenome), which is derived from relatively recent interspecific hybridizations between three different diploid species. In such polyploids, gene duplication alters the transcriptional landscape (4) by providing additional flexibility to adapt and evolve new patterns of gene expression for homoeologous gene copies (5). This flexibility has been suggested to be an important mecha-nism for controlling adaptive traits (6, 7)—for example, through neofunctionalization of duplicated genes (8) or tissue-specific expression (9). However, despite the likely importance of polyploidy in affecting gene expression, we have a limited understanding of the extent to which homoeologs resemble or differ from each other in their expression patterns, the spatiotemporal dynamics of these relationships, and how epi-static interactions between individual homoe-ologs affect biological traits. The new genomic resources available for wheat (10), along with its meiotic stability (11) and syntenic gene order (12), make it a particularly informative system for gaining insight into the effects of recent polyploidy on gene expression.

In this study, we leveraged available RNA sequencing (RNA-seq) data (529 samples from 28 studies) and added 321 samples to explore global gene expression in hexaploid wheat across a diverse range of tissues, developmental stages, cultivars, and environmental conditions (13). We organized these sets of RNA-seq samples into partially overlapping datasets from (i) a single developmental time course experiment (n = 209 samples), (ii) the reference accession Chinese Spring (CS) under nonstress conditions (n = 123 samples), (iii) four main tissue types under nonstress conditions (n = 537 samples), and (iv) seedling samples from abiotic (n = 50) and biotic (n = 163) stress experiments including controls (table S1). These datasets, alongside a complete and annotated genome and transcriptome (10), provide an opportunity to conduct homoeolog-specific transcriptome profiling and to generate gene regulatory networks to better understand the spatiotemporal coordination of individual homoeologs underlying trait biology on a genome-wide scale.

A developmental gene expression atlas in polyploid wheat

We first assessed expression patterns through a developmental time course of the commercial wheat cultivar Azhurnaya, including 209 RNA-seq samples representing 22 tissue types from grain, root, leaf, and spike samples across mul-tiple time points (Fig. 1). We quantified expres-sion using pseudoalignment of RNA-seq reads to the RefSeqv1.0 transcriptome, as implemented in kallisto (14), which accurately quantifies reads in a homoeolog-specific manner in polyploid wheat (13, 15) (figs. S1 and S2). We found evi-dence of expression for 83,741 (75.6% of 110,790) high-confidence genes, on the basis of expression of >0.5 transcripts per million (TPM) in at least one of the 22 tissue types, and we conducted complexity (table S2) and differential expression analyses (fig. S3). Tissue type distinguished sam-ples across development (fig. S4) (13), consistent with observations in other plant and animal spe-cies (16, 17). Within similar tissue types, subgenome of origin also influenced expression patterns, consistent with previous results in wheat grain samples (18). This gene expression atlas provides a valuable resource for breeders and researchers to query for and analyze their genes of interest through www.wheat-expression.com (15) and the Wheat eFP Browser at http://bar.utoronto.ca/ efp_wheat/cgi-bin/efpWeb.cgi (fig. S5) (19).

Homoeolog expression patterns

In polyploid wheat, quantitative variation for many agronomic traits is modulated by genetic interactions between multiple sets of homoeo-logs in the A, B, and D subgenomes (20). These

[1]John Innes Centre, Norwich Research Park, NR4 7UH Norwich, UK. [2]Plant Genome and Systems Biology, Helmholtz Center Munich, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany. [3]Earlham Institute, Norwich Research Park, NR4 7UZ Norwich, UK. [4]Bayer Crop Science, Innovation Center, Technologiepark 38, 9052 Zwijnaarde, Belgium. [5]Department of Cell and Systems Biology, Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada. [6]Saskatoon Research and Development Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada. [7]Crop Development Centre, University of Saskatchewan, Agriculture Building, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada. [8]Institut of Plant Sciences Paris-Saclay (IPS2), UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405 Orsay, France. [9]GDEC, INRA, UCA, 5 Chemin de Beaulieu, Clermont-Ferrand 63039, France. [10]Aquatic and Crop Resource Development, National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada. [11]INRA, 2 rue Gaston Crémieux, Evry 9057, France. [12]School of Life Sciences Weihenstephan, Technical University Munich, Munich, Germany. [13]IWGSC, 5207 Wyoming Road, Bethesda, MD 20816, USA. [14]School of BioSciences, University of Melbourne, AgriBio, La Trobe University, and School of Veterinary and Life Sciences, Murdoch University, 90 South Street, Perth, WA 6150, Australia. [15]Plant Sciences and Landscape Architecture, University of Maryland, 4291 Field House Drive, College Park, MD 20742, USA. [16]Global Institute for Food Security, University of Saskatchewan, 110 Gymnasium Place, Saskatoon, SK S7N 4J8, Canada. [17]Molecular Genetics, IPK Gatersleben, Corrensstrasse 3, 06466 Gatersleben, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: cristobal.uauy@jic.ac.uk (C.U.); philippa.borrill@jic.ac.uk (P.B.) ‡IWGSC collaborators and affiliations are listed in the supplementary materials. §Present address: Institute for Computational Biology, Faculty of Biology, University of Bielefeld, 33501 Bielefeld, Germany.

interactions range from buffering effects ob-served when gene homoeologs are function-ally redundant (21) to dominance effects where variation in a single homoeolog can lead to dominant phenotypes (22). Understanding how these interactions influence gene expression will help inform strategies to improve crops by target-ing and manipulating individual or multiple homoeologs to quantitatively modulate trait responses (20).

To determine patterns of homoeolog expression, we analyzed 123 RNA-seq samples representing 15 tissues under nonstress conditions (table S1) from CS. This was the same accession used to generate the reference genome (10); thus, cultivar-specific polymorphisms were ex-cluded from our analysis. We found evidence of expression for 82,567 (74.5%) high-confidence genes, consistent with the developmental time course of the cultivar Azhurnaya. We focused on 53,259 genes that had a 1:1:1 correspon-dence across the three homoeologous subgenomes, referred to as triads, and a summed expression of >0.5 TPM across the triad (64.5% of expressed genes, 96.1% of all triads; table S3). The majority of these expressed triads (94.3%) were in an-cestral (i.e., syntenic) physical positions in at least two of the three subgenomes [50,238 genes corresponding to 16,746 syntenic triads (10)], whereas 5.7% (1007 triads) had all genes in non-syntenic positions and are thus referred to as nonsyntenic triads. For each of the 17,753 expressed triads, we standardized the relative expression of the A, B, and D subgenome homoeologs (fig. S6) so that the sum was 1.0 in each individ-ual tissue. In this way, the relative abundance of homoeolog expression is comparable within triads, as well as across tissues, allowing the study of homoeolog expression bias (23).

We performed a global analysis combining data across all 15 tissues and focused on the 16,746 syntenic triads, but we discuss patterns in nonsyntenic triads where relevant. We found that the D subgenome had a subtly yet significantly higher relative abundance (33.65%) than the B (33.29%) and A (33.06%) subgenomes (Kruskal-Wallis $P < 0.001$; tables S4 and S5). The homoeolog ex-pression bias of the D subgenome is unlikely to reflect technical issues (fig. S2) and was found in 11 of the 15 tissues, was consistent across multiple expression abundance cutoffs, and was also significant in the developmental time course (in all 22 tissues) of the cultivar Azhurnaya [figs. S7 to S9 and table S5 (13)]. This effect, however, is subtle when compared with the homoeolog expression bias observed in evolutionarily older polyploid crops such as cotton, in which genome doubling occurred at an earlier time point (24).

The relative expression of each homoeolog determined a triad's position in the ternary plot for the global analysis (Fig. 2A), as well as for analyses of individual tissues (figs. S6 and S10). From these plots, we defined seven homoeolog expression bias categories (13): a balanced category, with similar relative abundance of transcripts from the three homoeologs, and six homoeolog-dominant or homoeolog-suppressed categories, classified on the basis of the higher or lower abundance of transcripts from a single homoeo-log with respect to those from the other two (Fig. 2A). Most syntenic triads (72.5%) were as-signed to the balanced category within each tis-sue, with balanced triads ranging from 62.6% in the stigma and ovary to 78.9% in roots (Fig. 2B and table S6). Triads with single-homoeolog dominance were infrequent (7.1%; range among tissues, 4.7 to 11.3%), whereas syntenic triads classified as single-homoeolog–suppressed were more common (20.5%; range, 16.3 to 27.1%; Fig. 2B). These patterns shifted significantly in the 1007 nonsyntenic triads, which had fewer balanced

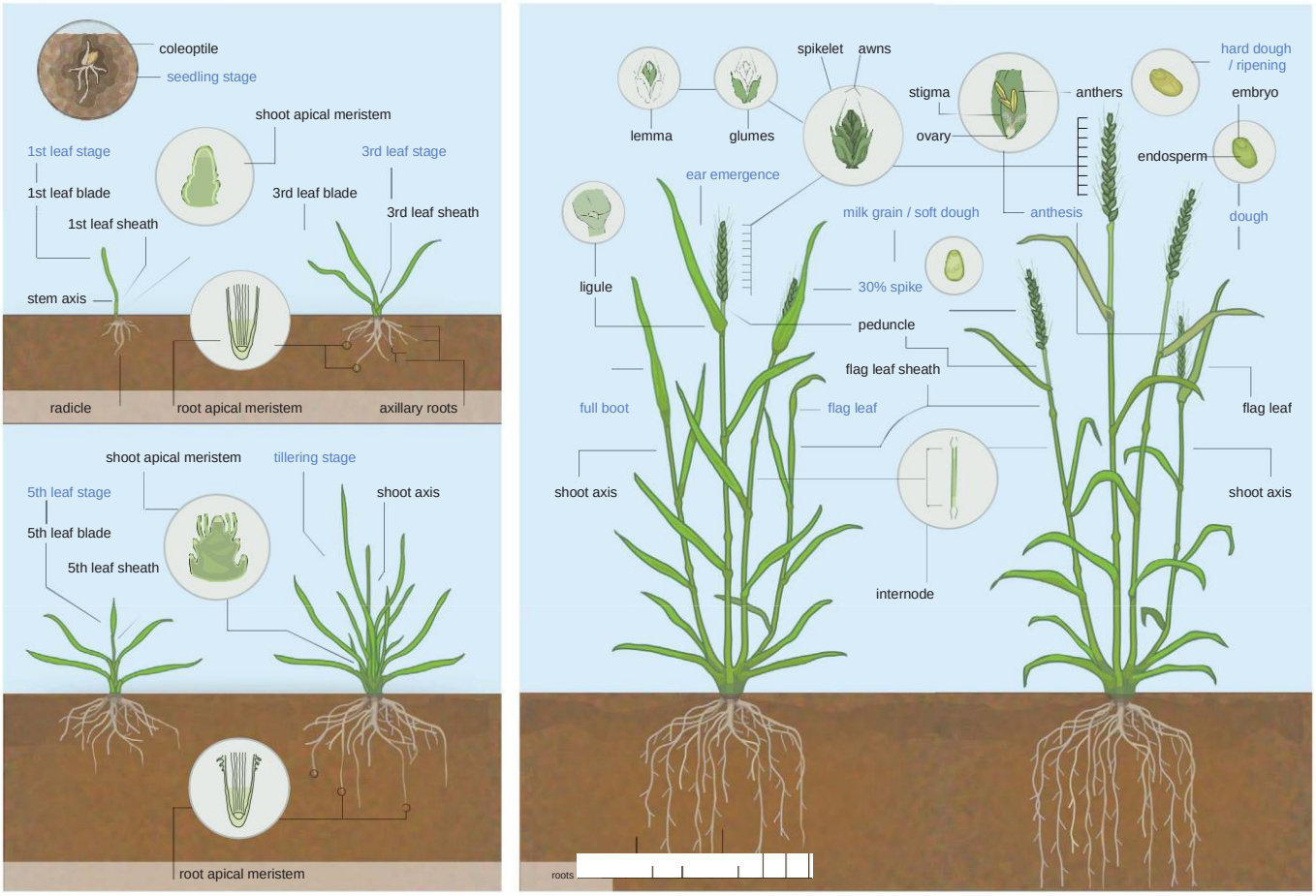


Fig. 1. Developmental time course of bread wheat. Shown is a schematic overview of tissues sampled for the RNA-seq expression atlas across multiple growth stages (labeled in blue). Details of all samples are provided in table S1.

triads (58.9%) and a higher proportion of dominant (14.5%) and suppressed (26.6%) triads across tissues (c^2 $P < 0.001$; tables S7 and S8). Across tissues, no differences were observed in the frequency of single-homoeolog dominance between subgenomes (tables S6 and S7). However, across all 15 tissues, D-homoeolog suppression was significantly less frequent (5.7%) than either A- or B-homoeolog suppression (7.5 and 7.2%, respectively; Kruskal-Wallis $P < 0.05$), and this pattern was also observed in nonsyntenic triads and the developmental time course (tables S6 to S8). This pattern explains in part the subtle homoeolog expression bias observed for the D subgenome relative to those observed for the

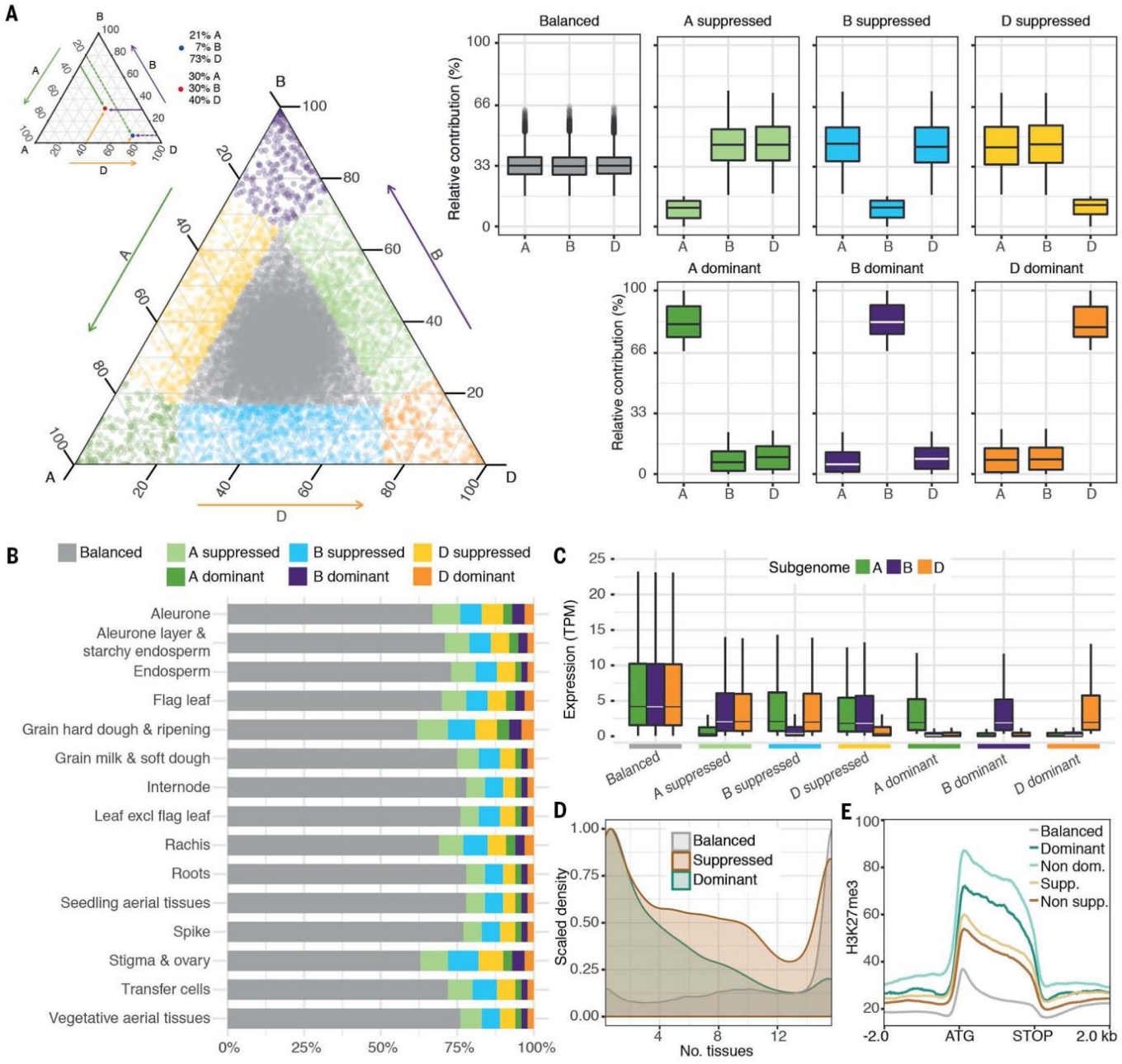


Fig. 2. Homoeolog expression bias in syntenic homoeolog triads.
(A) Ternary plot showing relative expression abundance of 16,746 syntenic triads (50,238 genes) in hexaploid wheat in the combined analysis of 15 tissues from Chinese Spring. Each circle represents a gene triad with an A, B, and D coordinate consisting of the relative contribution of each homoeolog to the overall triad expression (an example is shown on the top left) . Triads in vertices correspond to single - subgenome – dominant categories, whereas triads close to edges and between vertices correspond to suppressed categories. Balanced triads are shown in gray. Box plots indicate the relative contribution of each subgenome based on triad assignment to the seven categories.

A- and B-subgenome homoeologs. This observa-tion is consistent with the lower distribution of repressive H3K27me3 (histone H3 lysine 27 trimethylation) histone marks across the gene body of D-subgenome homoeologs compared with those of the A- and B-subgenome homoeologs (fig. S11).

Genes from syntenic triads in the balanced category were expressed across a wider range of tissues and had higher absolute transcript abundance, on a per-subgenome basis (12.2 tissues; median, 4.03 TPM), than genes in the suppressed (9.1 tissues; median, 1.51 TPM) or dominant (6.9 tissues; median, 0.57 TPM) categories (two-sample Kolmogorov–Smirnov test, $P < 0.001$; Fig. 2, C and D, and tables S9 and S10). The ab-solute transcript abundance data show that dominant triads are not the result of an overall increase in expression of a single homoeolog, but rather result from the relatively lower ex-pression of the two other homoeologs.

To determine if the differences among homoeologs are a consequence of polyploidization, we analyzed RNA-seq data from diploid and tetraploid progenitor species and newly created synthetic hexaploid wheat (SHW) lines (25). We found that 67.5% of nonbalanced triads in modern-day wheat have a different homoeolog expression bias category than that observed in SHW, with all three subgenomes being equally affected (table S11 and figs. S12 to S14). Likewise, 47.1% of nonbalanced triads in SHW are in a different category than would be expected on the basis of the progenitor species, with D-subgenome homoeologs most strongly influenced (13) (table S12 and fig. S15). These results suggest that the polyploid context and the polyploidization process itself affect the relative expression of homoeologs compared with the baseline expression in the progenitor species (13), which has also been observed during the evolution of polyploid cotton (26) and monkeyflower (27).

We hypothesized that epigenetic mechanisms might be associated with differences in homoeolog expression patterns. To test this, we examined the associations of transposable elements (TEs), DNA methylation, and histone modifications with the relative expression of triads in leaves of CS. We found no clear relationship between the presence of TEs in promoter regions and altered expression patterns between ho-moeologs in dominant and suppressed triads (Tukey's Honestly Significant Difference $P > 0.6$; fig. S16 and table S13) (13). However, we identified significant differences in gene-body DNA methylation and histone modifications among homoeologs (13).

Gene-body CG methylation is widely conserved in angiosperms, although its functional significance is currently under debate (28, 29), given that two angiosperm species lack this epigenetic mark altogether (30). We found higher gene-body CG methylation in constitutively expressed triads than in more tissue-specific triads (balanced > suppressed > dominant; fig. S17). Within the nonbalanced triads, homoeologs with higher expression had higher CG methylation than their corresponding nondominant and suppressed homoeologs (Mann-Whitney $P < 0.001$; fig S17). These results are consistent with gene-body CG methylation associated with housekeeping genes and its suggested role in homoeostatic gene expression (29). Similarly, the more highly expressed homoeologs within nonbalanced tri-ads had higher active (H3K36me3 and H3K9ac; ac, acetylation) histone marks and lower re-pressive (H3K27me3) histone marks in the gene body (Mann-Whitney $P < 0.001$; fig. S11). For H3K27me3, these differences were not limited strictly to the gene body but extended into the upstream and downstream regions for both dominant and suppressed triads (Fig. 2E), consistent with the tight association of H3K27me3 with inactive gene promoters (31). These results suggest that epigenetic status in gene bodies, as well as upstream and downstream regions, is associated with homoeolog expression bias in polyploid wheat, consistent with results in monkeyflower showing changes in DNA methylation upon polyploidization (27).

Breeders rely on recombination to generate new combinations of haplotypes for improving cultivars. In wheat, chromosome position strong-ly influences recombination rates, with relatively low recombination rates in the interstitial and proximal regions (R2a, C, and R2b genomic com-partments) but markedly higher rates toward the distal ends of the chromosomes (R1 and R3 genomic compartments) (32). In our analyses, syntenic triads in the balanced category were overrepresented in the low-recombination re-gions (R2 and C), which have higher levels of ac-tive histone marks (H3K36me3 and H3K9ac) (10), consistent with the higher expression of balanced triads. Homoeolog-dominant and homoeolog-suppressed triads were overrepresented toward the high-recombination distal ends of chromo-somes (R1 and R3; c^2 $P < 0.001$; table S14), which have higher levels of repressive (H3K27me3) his-tone marks (10), consistent with the lower expres-sion of dominant and suppressed triads. This pattern was also observed in the developmental time course of the cultivar Azhurnaya. However, when comparing the CS and Azhurnaya cultivars (nine tissues in common), we found that 84.5% of genes in the R2 and C regions had the same expression category between cultivars, whereas only 72.2% of genes in the R1 and R3 regions did so (c^2 $P < 0.001$; table S15). These differences in homoeolog expression bias across cultivars have important implications for breeding because they suggest that through genetic crosses, breeders not only generate new combinations of haplo-types with differential expression of alleles, but also rearrange and select for homoeolog expres-sion bias between cultivars.

Variation of triad expression patterns

Polyploidy may confer phenotypic plasticity by allowing homoeologs to be expressed differently across tissues and/or environmental conditions (8). Our analyses above provide a static overview of the relative homoeolog expression bias in individual tissues. Therefore, we explored whether syntenic triads retain their homoeolog expression bias category across the 15 tissues (table S16). We found that 83.6% of balanced triads remained balanced in each of the 15 individ-ual tissues, whereas dominant and suppressed triads tended to be more variable across tissues, with only 73.4 and 62.2%, respectively, staying within their global dominance group across all 15 tissues (Fig. 3A). Dominant and suppressed triads shifted most often to adjacent categories (16 to 20%) in the ternary plots and in few cases (<3.0%) changed to opposite categories (fig. S18). These patterns were also observed in the devel-opmental time course (table S16). These data show that across tissues, triads most often re-mained consistent in their homoeolog expres-
sion bias classification, a phenomenon also seen across seven tissues in allotetraploid Tragopogon mirus (33).

To complement this analysis, we determined the variation in behavior of each triad within the ternary plot across the 15 tissues by calculating the mean distance between the triad's position in each tissue and its global average position (13) (fig. S19). This generated a distribution of mean distances (Fig. 3B); we focused on the 10% most stable triads (defined as those having the short-est mean distances across tissues) and the 10% most dynamic triads (largest mean distances) (Fig. 3, B to D). Stable triads were expressed more highly than dynamic triads (median, 8.2 versus 3.2 TPM; $P < 0.001$) and had a higher expression breadth, being expressed across almost all sam-ples, whereas dynamic triads were more tissue-specific ($P < 0.001$) (Fig. 3E and table S17). Stable triads were enriched for high-level gene ontology (GO-slim) terms associated with housekeeping processes (e.g. translation and cell cycle), whereas dynamic triads were enriched for defense and external stimuli responses and secondary meta-bolic processes, functions that more frequently determine differences in individual fitness (table S18) (6). In the global analysis, stable triads were significantly enriched for the balanced category (94.2%), whereas dynamic triads were almost equally spread between suppressed (37.9%), dom-inant (30.5%), and balanced (31.6%) categories (c^2 $P < 0.001$; Fig. 3F and table S19). This pat-tern is consistent with stable triads being most frequently located in proximal regions (C), whereas dynamic triads tend to locate in distal ends of chromosomes (R1 and R3) (c^2 $P < 0.001$; table S20). These results demonstrate expression asymmetry across wheat chromosomes, whereby the high-recombination distal ends of chromosomes have triads that exhibit higher homoeolog expression bias, are more dynamic across tissues, and have higher expression variation between cultivars than triads in the low-recombination proximal regions (Fig. 3G). This asymmetry is also re-flected in the contrasting distributions of histone marks and DNA methylation along chromosomes (10). The difference in epigenetic marks identi-fied in leaves makes it tempting to speculate that epigenetic marks may also be associated with triad expression variation across multiple tissues.

We next investigated if divergence in spatial expression patterns (as measured by the mean

distance statistic described above) was coupled with 5′ promoter and protein sequence diver-gence in syntenic triads (13). The 1.5–kilobase (kb) promoters of dynamic triads more frequently contained TEs (88.3 versus 79.2%), which were closer to the translation start site (1113 versus 1234 bp away) but shorter (median, 220 versus 259 bp), than those in stable triads, leading to equivalent TE densities (all comparisons, Kruskal-Wallis P < 0.001; fig. S20 and table S13). These closer, more frequent, and shorter TEs could potentially act as novel cis-regulatory elements (34) or influence epigenetic marks (35). These results indicate that the promoter TE landscape relates more closely to the variation in the rela-tive expression of homoeologs across tissues than to a ubiquitous effect across all tissues (table S13). Although only subtle differences in sequence iden-tity were found between stable and dynamic triads (85.5 versus 85.0%; P = 0.045) (Fig. 3H and table S21), dynamic triads had fewer con-served transcription factor (TF) binding site motifs across the three homoeologs (37% fewer; P < 0.001; fig. S21). Across coding sequences, we showed a stepwise decrease in conservation of both the nucleotide and protein identities from stable (average, 97.2% coding sequence and 97.3% protein) to dynamic (95.0 and 93.4%) triads (both P < 0.001; table S21). We compared nonsynony-mous (K_a) with synonymous (K_s) substitution

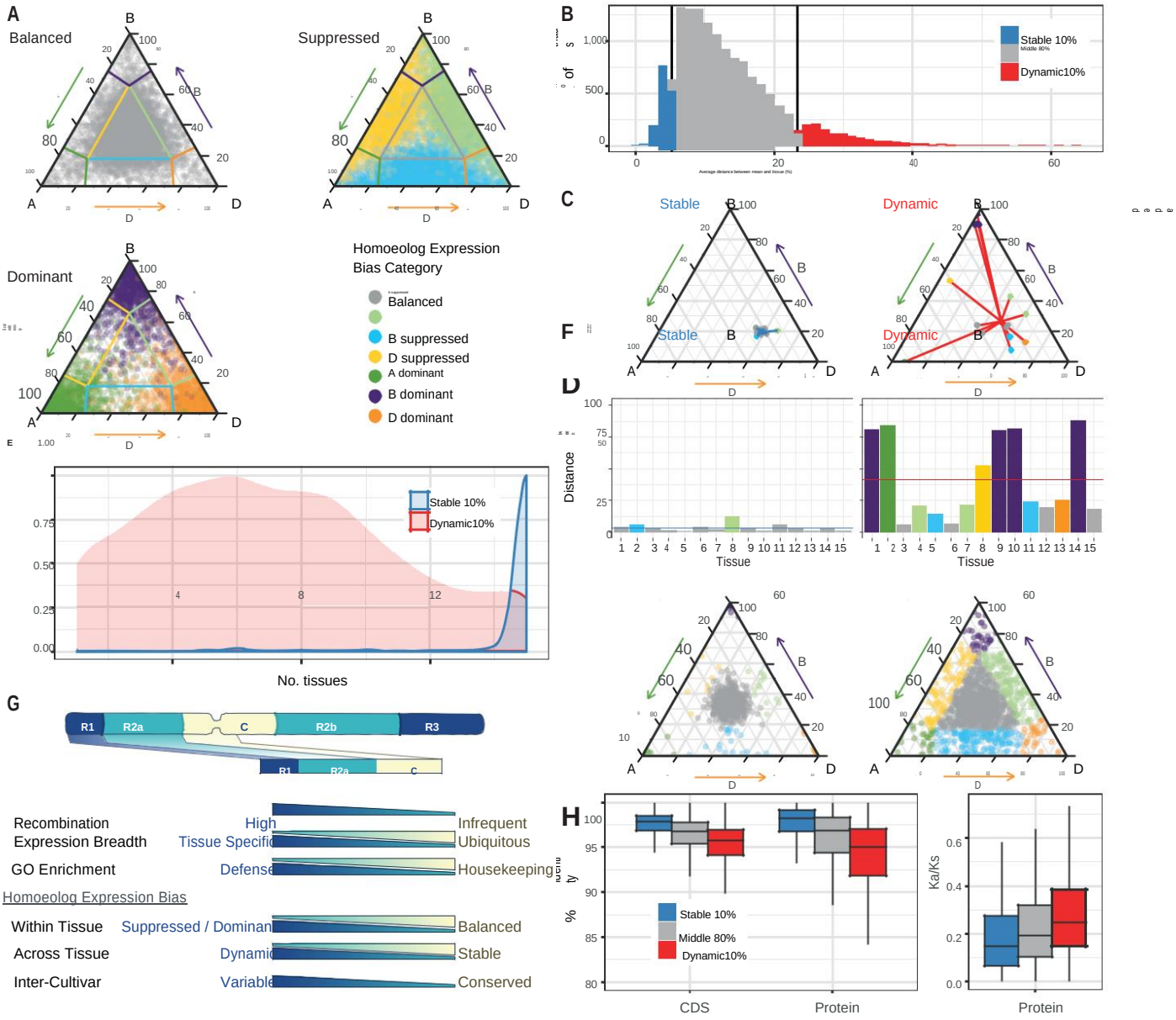


Fig. 3. Variation of triad expression patterns. (A) Variation of balanced, dominant, and suppressed triads (assigned on the basis of global analysis) across 15 tissues. (B) Distribution of mean distance of triad variation across 15 tissues for 14,258 triads expressed in at least six tissues. The 10% most stable (blue) and 10% most dynamic (red) triads were defined. (C) Ternary plots of representative stable and dynamic triads and (D) bar graphs of the distance between the triad position in the 15 individual tissues and the triad global average position (horizontal line). Color-coding is as in (A). (E) Number of tissues in which stable (blue) and dynamic (red) genes are expressed (table S36). (F) Homoeolog expression bias classification of stable and dynamic triads in global analysis. (G) Schematic representation of a wheat chromosome based on genomic compartments and features associated with distal (R1 and R3) and interstitial and proximal (R2 and C) regions. (H) Box plots of percent coding and protein sequence identity (left) and K_a/K_s ratio (right) for stable 10% (blue), middle 80% (gray), and dynamic 10% (red) triads.

rates between homoeologs and observed that dynamic triads had significantly higher K_a/K_S than stable triads (0.33 versus 0.21; Mann Whitney $P < 0.001$; Fig. 3H and table S22). This higher ratio suggests that triads with greater divergence in spatial expression patterns are under more relaxed selection pressure, as seen for duplicated genes in humans (9), but not in soybean (36) and carp (37). This conclusion is supported by the ob-servation that nonsyntenic triads, which had greater expression divergence (10.5% larger mean distance; Mann-Whitney $P < 0.001$), also had significantly higher K_a/K_S (0.42; Mann-Whitney $P < 0.001$) compared with syntenic triads (table S22). The above relationships were consistent when using different percentage cutoffs to define stable and dynamic triads (5 and 25%), as well as in the developmental time course of the cultivar Azhurnaya (tables S21 and S22). These results show positive coupling of divergence in spatial expression patterns with divergence in TE and cis-regulatory elements in promoters and se-quence divergence in coding sequence among wheat homoeologs. It is possible that diver-gence in spatial expression patterns, alongside relaxation of selection pressure, can lead to func-tional innovation through homoeolog neo- or subfunctionalization.

Coordinated expression of homoeolog triads

Our analyses provide a framework to describe the relative expression of individual homoeologs between discrete triads in space and time. To understand how this coordination of homoeo-log spatiotemporal expression may influence biological processes, we developed a series of coexpression networks to provide insight into tissue-specific developmental and stress-related processes.

We constructed four separate tissue-specific coexpression networks from nonstress RNA-seq samples from grain (n = 119 samples), leaf (n = 245), root (n = 45), and spike (n = 128), using all genes expressed at more than 0.5 TPM in the given tissue (13). These networks were com-posed of 51 to 78 modules and contained 42.3 to 88.0% of all expressed genes in each tissue (fig. S22, table S23, and data S1). We found that across all tissue networks, homoeologs from 37.4% of the syntenic triads were in the same coexpression module, suggesting a highly coor-dinated expression pattern for these triads (c^2 $P < 0.0001$ with respect to random triads; table S24). However, the majority of triads (62.6%) had at least one homoeolog outside the same module.

To quantify whether homoeologs outside the module had similar or divergent expression pat-terns, we calculated a threshold based on the pairwise distance between homoeologs (13). We found that 29.6% of syntenic triads had a diver-gent pattern, wherein the expression of at least one homoeolog exceeded the distance threshold in the tissue network (Fig. 4A and fig. S23). Conversely, 33% of triads had a similar pattern, wherein all pairwise distances between homoeo-logs were lower than the threshold, suggesting a

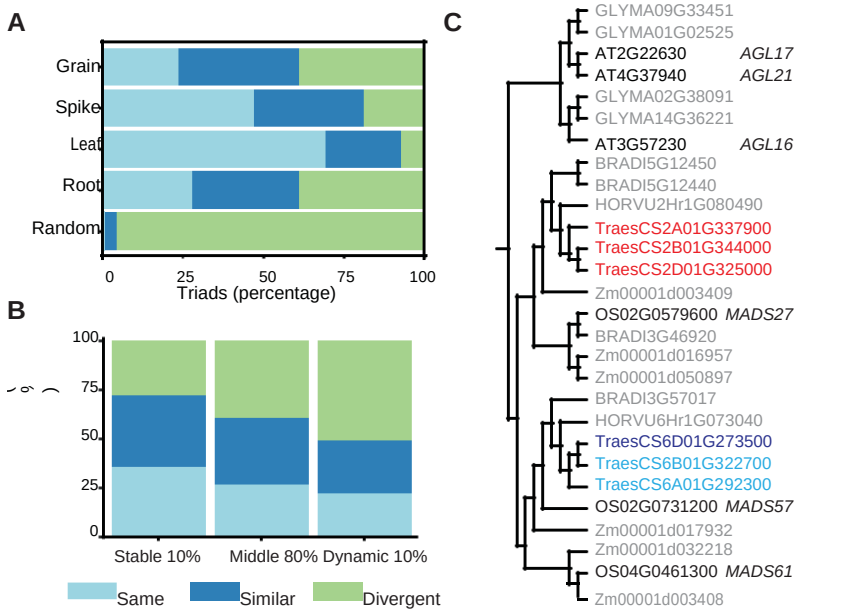


Fig. 4. Homoeolog coexpression patterns in tissue networks. (A) Triad assignment to same, similar, and divergent modules in tissue coexpression networks. (B) Stable, middle, and dynamic triad assignment to same, similar, and divergent modules in the root network. (C) Neighbor-joining phylogenetic tree of homologs for the Arabidopsis MADS_II gene AGL21. Wheat orthologs from chromosome group 2 are assigned to the root-specific module 61 (red), whereas chromosome group 6 orthologs are assigned to modules 1 and 13 (blue and purple).

subtler variation in a single homoeolog. These values showed significant variation between tis-sue networks, ranging from 7% divergent triads in the leaf network to more than 38% divergent triads in the root and grain networks (Fig. 4A). Nonsyntenic triads had a higher proportion of divergent triads in all tissue networks compared with syntenic triads (mean, 35.1 versus 29.6%; c^2 $P < 0.001$; table S24). Using the same crite-rion as before (triad mean distance between tis-sues and global average position), we identified the 10% most stable and dynamic syntenic triads for each tissue-specific network. We found that dynamic triads were more frequently in diver-gent modules than stable triads for all four tis-sue networks ($P < 0.001$; Fig. 4B and fig. S24). These results are consistent with the homoeo-log expression bias analyses and support the idea that although many triads are expressed in a coordinated spatiotemporal pattern (with the same or similar profile), almost 30% of syntenic and 35% of nonsyntenic triads have a divergent expression profile. Transcriptional divergence occurs both immediately upon poly-ploidization and after polyploidization (figs. S12 to S14) and may represent initial steps toward neo- or subfunctionalization of wheat homoeologs.

Exploiting development and stress networks for biological discovery

To explore the potential for biological discovery, we first compared modules between networks to identify tissue-specific gene networks. Across the four networks, 73.2% of modules had significant overlap (Fisher's exact test, $P < 0.05$) with mod-ules in all four networks, with the root having the fewest conserved modules (61.1%) and the spike having the most (86.2%) (data S2). In the root, there were three modules that were not found in any other tissue, with the largest of these (root module 61; 82 genes) enriched for root-related plant ontology (PO) terms (e.g., root procambium, $P = 3.3 \times 10^{-5}$, and central root cap of primary root, $P = 4.5 \times 10^{-5}$; table S25). We hypothesized that genes encoding TFs controlling processes re-lated to these PO terms would also be coexpressed within module 61. We found that four of the 10 genes encoding TFs in this root-specific module had known functions related to root development in Arabidopsis or rice (38, 39) (table S26). Three of these TFs belonged to one homoeolog triad in the MADS_II family, and one of their Arabidopsis orthologs (AGL21) has been shown to regulate lateral root development through auxin accu-mulation (39). To understand the target genes of these TFs in wheat, we conducted a comple-mentary network analysis using genie3, which predicted target genes of TFs across all 850 sam-ples (13). Target genes of the three TFs were enriched for cell wall processes and lignifica-tion, consistent with their putative role in the differentiation zone where lateral roots emerge (tables S27 and S28). Closely related paralogs on chromosome group 6 in wheat were not located in root module 61; rather, they were in modules 1 and 13 (Fig. 4C). These modules were conserved in all other tissue networks, implying a more

general function for genes within them. Supporting this hypothesis, the rice ortholog of the chromosome 6 paralogs (OsMADS57) has been shown to play a role in tillering (40).

A key challenge for wheat breeding is the selection of cultivars with tolerance to multiple stresses. Therefore, we focused on stress responses in seedlings and young vegetative plants, for which 10 independent studies with 12 distinct abiotic and disease stresses were available (table S1). We constructed gene coexpression networks for abiotic and disease stresses separately, including control samples from the same studies to allow for links between disease status and gene expression (13). We integrated the two networks to identify modules that might be common to both abiotic and disease responses. We found 84 pairs of modules between the two networks that had significantly overlapping gene content and were significantly correlated with both an abiotic and a disease stress (tables S29 to S31). The most significant overlap was between disease module 12 and abiotic module 2 ($P = 1.3 \times 10^{-94}$), which shared 355 genes (Fig. 5A). These two modules had similar enrichment for GO-slim terms relating to signal transduction and response to stimulus (table S32), suggesting that they might perform similar biological functions.

Among the 355 shared genes, there were 16 encoding TFs, six of which have orthologs in rice or Arabidopsis with proven roles in abiotic or disease stress, and a further three have orthologs differentially expressed during stress in these species (table S33). Furthermore, on the basis of the genie3 analysis, 11 of the 16 TFs have targets that are enriched in stress responses, and seven have targets that are enriched simultaneously in biotic and abiotic stress responses (Fig. 5B). Of the genes encoding these TFs, two homoeologs stood out as potential common regulatory components of abiotic and disease response: TraesCS5A01G237900 and TraesCS5B01G236400, which encode heat shock factor (HSF) TFs. These two HSF TF–encoding genes were in the top 10 most central genes within disease module 12 (table S34), as measured by intramodule connectivity, a value strongly correlated with the influence of a gene on a phenotype (41). The 387 predicted targets of the TFs encoded by these two genes were frequently allocated to module 12 of the disease network (39.5%) and module 2 of the abiotic stress network (28.0%) (table S35). The Arabidopsis ortholog of these genes, TBF1, was originally identified for its role in pathogen defense response (42) and has been shown to play a key role in the transition from growth to defense (43), while also positively regulating acquired thermotolerance (44). Recently, a "TBF1 cassette" including the promoter and 5′ leader region of TBF1 was used to engineer broad-spectrum disease resistance in both Arabidopsis and rice without a fitness cost (45). The fact that Arabidopsis TBF1 is functional in rice suggests that this regulatory mechanism may be conserved across species, making the wheat orthologs identified here promising targets for further studies. These and other highly connected genes (table S34) are strong candidates for controlling stress responses, and the functions of their orthologs support this hypothesis. These results demonstrate the power of the datasets and show that integrating gene networks from wheat, alongside phylogenetic relationships and knowledge of biological function in model species, can help identify candidate genes for further study in wheat.

Concluding statement

This study provides detailed insight into the spatiotemporal transcriptional landscape of polyploid wheat. We find evidence that the differences in relative expression among homoeologs observed in modern-day wheat may have been established both upon the polyploidization of wheat itself and during the subsequent 10,000 years of polyploidy; these differences may have been determined through epigenetic changes affecting both DNA methylation and histone modifications. We identified asymmetries along wheat chromosomes for a series of features relating to homoeolog expression bias with important implications for breeding. Our work provides a framework for the generation of hypotheses about biological function in polyploid wheat, which can now be experimentally tested using recent developments in sequenced mutant populations (46) and genome editing approaches (47). Ultimately, this knowledge will help researchers and breeders modulate allelic variation across homoeologs to improve quantitative traits in polyploid wheat. This is an urgent task for achieving global food security, given that wheat provides more than 20% of the protein and caloric intake (48) of humans.

Materials and methods

RNA-seq samples

We included 246 samples previously described (15) and complemented this with 283 RNA-seq samples which were deposited in the Short Reads Archive (SRA) between August 2015 and January 2017. An additional 321 RNA-seq samples from six studies were used for this analysis and are detailed in the supplementary materials (13).

Mapping of RNA-seq reads to reference

For all 850 samples, metadata was assigned as described in (15), with high and low-level factors for tissue, age, variety, and stress. Due to the relatively large number of low-level tissues (59), we

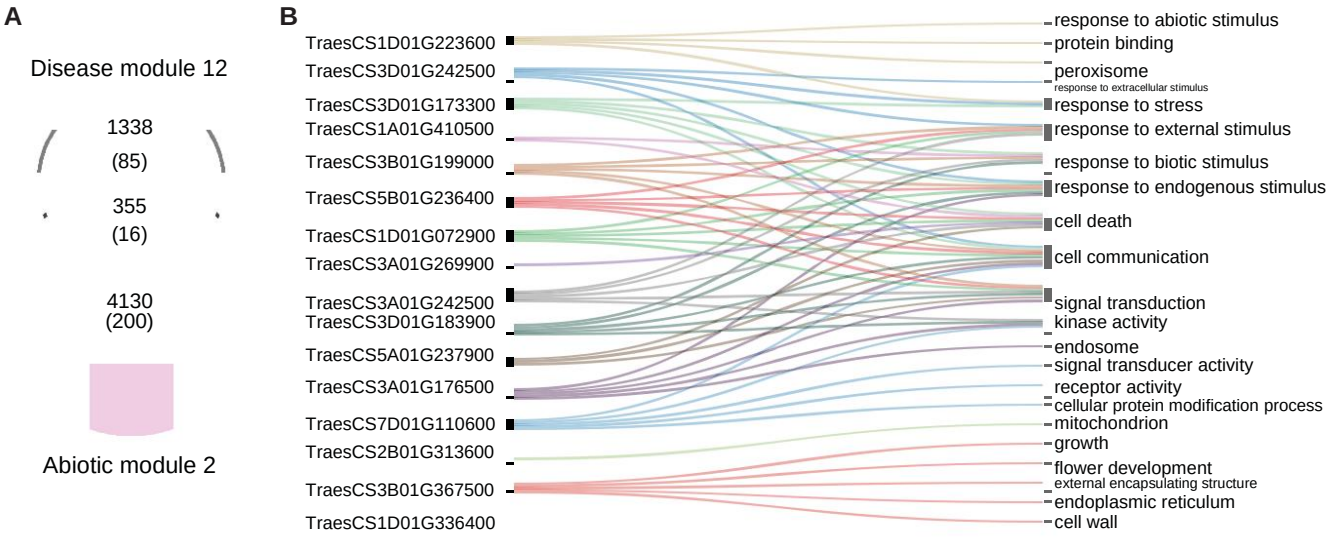


Fig. 5. Overlapping modules within abiotic and disease stress networks. (A) Number of genes in abiotic module 2 and disease module 12 and the overlap between modules. The number of transcription factors is indicated in parentheses. (B) Transcription factors found in both abiotic 2 and disease 12 (left) and the top five enriched GO terms of their targets, as identified by genie3 (right).

further defined an intermediate level of tissues comprising 32 factors (average 26.5; median 12 replicates per factor) which was used for this study. We also assigned an intermediate level of stress comprising 15 factors (average 14.5; median 6 replicates per factor). We used kallisto v0.42.3 (14) to map the 850 RNA-seq samples to the Chinese Spring RefSeqv1.0+UTR tran-scriptome reference. We used default param-eters previously shown to result in accurate homoeolog-specific read mapping in polyploid wheat (15) (fig. S1). We summarized expression levels from the transcript level to the gene level using tximport v1.2.0. We established the crite-ria that at least 1% of samples for a given gene all
required to have expression values over 0.5 TPM for that gene to be considered expressed (initial 850 filter).

To confirm that kallisto enables homoeolog specific mapping (15) we analyzed expression of HC genes expressed >0.5 TPM in nulli-tetrasomic wheat lines from the publicly available study SRP028357 (49). The nulli-tetrasomic lines were missing an entire chromosome (1A, 1B, or 1D) which was replaced by a duplication of another homoeologous chromosome, e.g. Nulli1ATetra1B has 0 copies of 1A, 2 copies of 1B and 1 copy of 1D. We determined stringent homoeolog-specific mapping using a series of criteria detailed in the supplementary materials (13).

Analyses of expressed genes

Starting from the subset of genes considered expressed using the initial 850 filter criterion,
we determined genes which were expressed in at least one tissue within the Azhurnaya devel-opmental time course (209 samples; 22 inter-mediate tissues) and Chinese Spring no stress (123 samples; 15 intermediate tissues) datasets. For this analysis, we first calculated the average
TPM expression of each gene in each of the in-termediate tissue types (average expression per tissue). The number of samples that went into generating this average expression per tissue
value varied for each intermediate tissue and
are available in table S1. We considered a gene expressed when its average expression per tissue
was > 0.5 TPM in at least one intermediate tissue. For both datasets we focused on HC gene models (10). Whilst expression data was also assessed for LC genes, we excluded these from the main analysis to avoid confounding effects from pseudogenes and low-quality gene models. Through this analysis we found evidence of ex-pression for 83,741 (75.6%) HC genes in Azhurnaya
and 82,567 (74.5%) HC genes in Chinese Spring.

Using the average expression per tissue values, we also determined the global expression of each gene across all tissues in which it was expressed (based on the >0.5 TPM criteria in the tissue). This generated an average value across tissues, rather than a geometric mean across all sam-ples, to account for the variation in the number of samples per tissue. It also excludes tissues in which a gene is not expressed. This average across expressed tissues is referred to as either the "global analysis" or the "combined analysis (all tissues)" across the main text and in the supplementary materials and tables.

Relative expression levels of the A, B, and D subgenome homoeologs across triads

The analysis focused exclusively on the gene triads which had a 1:1:1 correspondence across the three homoeologous subgenomes, including 17,400 syntenic and 1074 nonsyntenic triads (total of 18,474 triads or 55,422 genes). Start-ing from the subset of genes considered ex-pressed using the initial 850 filter criterion, we defined a triad as expressed when the sum of the
A, B, and D subgenome homoeologs was > 0.5 TPM. This allowed us to include triads in which, for example, only a single homoeolog was expressed, and which could later be classified as a dominant triad. Using this criterion, we defined a total of 53,259 genes (17,753 triads) which were con-sidered expressed (table S3). To standardize the relative expression of each homoeolog across the triad, we normalized the absolute TPM for each gene within the triad as follows

$$\text{expression}_A = \frac{\text{TPM(A)}}{\text{TPM(A)} + \text{TPM(B)} + \text{TPM(D)}}$$

$$\text{expression}_B = \frac{\text{TPM(B)}}{\text{TPM(A)} + \text{TPM(B)} + \text{TPM(D)}}$$

$$\text{expression}_D = \frac{\text{TPM(D)}}{\text{TPM(A)} + \text{TPM(B)} + \text{TPM(D)}}$$

where A, B, and D represent the gene corre-sponding to the A, B, and D homoeologs in the triad. The normalized expression was calcu-lated for each one of the intermediate tissues
and for the average across all expressed tissues ("combined analysis" as described previously).
Fig. S6 shows an example of these calculations for the roots and the combined analysis across three triads. The values of the relative contri-butions of each subgenome per triad were used
to plot the ternary diagrams using the R package ggtern (50).

Definition of homoeolog expression bias categories

The ideal normalized expression bias for the sev-en categories was defined as shown in table S37.

We calculated the Euclidean distance (rdist function from R 3.3.2) from the observed nor-malized expression of each triad to each of the seven ideal categories listed above. We assigned the homoeolog expression bias category for each triad by selecting the shortest distance. This was done for each of the intermediate tissue as well
as for the average across all expressed tissues (combined analysis).

Analysis of the effects of polyploidy on homoeolog expression bias

We used RNA-seq data (25) which consisted of two datasets based on RNA-seq samples from the youngest leaf at fifth leaf stage. Dataset 1 (SHW1) included samples from tetraploid (BBAA) Triticum turgidum ssp. turgidum wheat acces-sion AS2255, diploid Ae. tauschii (DD) accession
AS60, and the synthetic hexaploid wheat (SHW1; BBAADD) resulting from the cross between the tetraploid and Ae. tauschii accessions. Dataset 2 (SHW2) consisted of tetraploid T. turgidum ssp. durum cv Langdon (BBAA), the same diploid Ae. tauschii (DD) accession AS60, and an indepen-dent synthetic hexaploid wheat (SHW2) derived from Langdon x AS60 (BBAADD). Note that AS2255 and Langdon are both T. turgidum ssp., but are defined as different subspecies based primarily on morphological features. These ex-periments recreate the polyploidization events that gave rise to modern bread wheat and the resulting SHW has the same genome compo-sition as the CS and Azhurnaya datasets exam-ined in this study.

We analyzed the RNA-seq from both data-sets by mapping reads to the CS RefSeqv1.0 transcriptome using the same bioinformatics pipeline as before (see "Mapping of RNA-seq reads to reference" section). However, for the tetraploid datasets we used only the A and B subgenome transcripts as a reference, for the dip-loid D genome datasets we used only D subgenome transcripts, and for the SHW datasets we used the complete RefSeqv1.0 transcriptome as the reference, as in CS and Azhurnaya. To generate the expected hexaploid wheat transcriptome based on progenitor species we weighted the TPM values from the tetraploid by 2/3 and the

AS60 TPM values by 1/3 to maintain a total TPM of 10^6 in the combined dataset. The in-silico hexaploid wheat generated from the weighted tetraploid and diploid TPM values (referred to hereafter as the "expected" in-silico dataset) allows the direct comparison with the observed
TPM values in SHW. We defined the seven ho-moeolog expression bias categories for both the expected in-silico and the observed SHW tran-scriptomes using the same methods as for CS and Azhurnaya and compared the classification of triads between the observed and expected data-sets (table S12). We next compared classifications to modern-day bread wheat CS and Azhurnaya. To enable a meaningful comparison across sim-ilar tissues from the Hao et al study (25) we used nine samples from the PAMP Triggered Immune Response dataset from CS and six samples from the Azhurnaya dataset (table S1). As before, we defined the seven homoeolog expression cat-egories for the defined CS and Azhurnaya data-sets and compared them with the SHW and the in-silico classifications (table S11).

DNA methylation plant material and library preparation

Plants were grown as described in the Chinese Spring tissues study. The frozen leaves from the five samples at 3-leaf stage (Zadok stage 13) were ground and divided as input for the prep-aration of both RNA-seq libraries (detailed in Chinese Spring tissues study) and whole genome bisulfite sequencing (WGBS) libraries. These samples enabled direct comparisons between

the DNA methylation profile and homoeolog expression patterns in the same samples. WGBS libraries were constructed from purified nuclei prepared using the published methods (51). Input DNA was quantified using the Qubit high sensitivity DNA kit. A total of 500 ng of nuclear DNA was spiked with 270 pg of Lambda DNA to assess the conversion efficiency obtained using the EZ DNA Methylation-GoldTM Kit (Zymo research corp, Irvine, Ca, USA). WGBS libra-ries were prepared using the TruSeq DNA kit, (Illumina, Madison, WI) and 2 × 125 bp paired-end sequence reads was generated using the Illumina HiSeq 2500 v4 platform (Genome Quebec, Montreal, QC, Canada). The data was deposited as SRP133674.

DNA methylation data analysis

Sequence quality and adaptor removal was performed using Trim_galore_v0.4.1 (52). High quality paired-end sequence reads were aligned to the RefSeqv1.0 Chinese Spring genome using Bismark version 0.16.1 (53) ensuring the remov-al of duplicate reads and only retaining unique unambiguous alignments. The data were processed to exclude regions with low coverage using a binomial test. The methylation data was annotated using the gene feature coordinates provided by the RefSeqv1.0 Chinese Spring gene definitions. 5 kb flanking regions around the gene features were also extracted. The two flanking regions and the gene feature were each divided into 50 tiles (150 tiles in total) to summarize the observed methylation ratios. Data manipulation, statistical analysis and image generation were performed using the R language (54) utilizing the data.table (55), MethylKit v1.5.2 (56), genomation (57), and ggplot2 packages (58).

Comparison of RNA-seq sample classification with DNA methylation

For the five RNA-seq samples (from the same plants used for analyzing DNA methylation) we classified triads into the seven balanced, dominant, and suppressed categories using the same method as for previous analyses. We then classified homoeologs within dominant and suppressed triads into the “dominant” and “nondominant” homoeologs, and “suppressed” and “nonsup-pressed” homoeologs. For example, in an A dominant triad the A subgenome is classified as “dominant” and B and D subgenomes are classified as “nondominant” homoeologs. The DNA methylation patterns of genes in each of these categories were plotted using the methods described above (DNA methylation data analysis). Differences in DNA methylation levels between categories were tested pairwise using the non-parametric Mann-Whitney t test using the wilcox.test() in R (fig. S17).

Histone modification analysis

To study the role of histone modifications we carried out ChIP-seq for three active marks (H3K36me3, H3K9ac, and H3K4me3) and one repressive mark (H3K27me3) (deposited under SRA accession number SRP126222). We used Chinese Spring at 3-leaf stage, however RNA-seq data were not collected from these exact plants. To calculate the homoeolog expression bias we used Chinese Spring samples from a separate experiment (PAMP-triggered immune response study) in which the same tissue was collected at a similar stage (3-leaf stage). Whilst combining data from two separate experiments may introduce some noise into the analysis, the ChIP-seq and RNA-seq data are from similar tissues, at a similar growth stage, in the same wheat variety, and are thus highly comparable. Nevertheless, this confounding factor should be considered when interpreting these results. ChIP assays, DNA library preparation, and se-quencing were performed as in (10).

Histone data analysis

Raw FASTQ files were preprocessed with Trimmomatic v0.36 (59) to remove Illumina sequencing adapters, trim 5′ and 3′ ends with quality score below 5 (Phred+33) and discard reads shorter than 20 bp after trimming. Paired-ends reads were aligned against IWGSC RefSeq v1.0 assembly using bowtie2 v2.3.3 with –very-sensitive settings (60). Alignments with MAPQ < 10 were discarded and duplicate reads removed with Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Triad expression category was calculated using Chinese Spring samples from a separate experiment (PAMP-triggered immune response study) using the same method as in previous analyses. As in the DNA methylation analysis we classified homoeologs within dominant and suppressed triads into the “dominant” and “nondominant” homoeologs, and “suppressed” and “nonsuppressed” homoeologs.

We calculated meta-gene profiles for each category by computing the read density of each histone mark over different triads categories using Deeptools (61) computeMatrix scale-regions and plotted it with plotProfile. To make a statistical comparison, for each histone mark we scored the number of reads overlapping with gene bodies using bedtools coverage –counts (62). Only reads fully mapping within gene bodies were considered. To account for differ-ent gene size we divided the read counts over each gene by its length. The distributions of reads density over different triads categories were compared with a nonparametric t test (Mann-Whitney U-test) using the function wilcox. test in R (fig. S11).

Variation in homoeolog expression bias across tissues (stable and dynamic triads)

To define the variation in homoeolog expression bias of each triad across the intermediate tissues we calculated the Euclidean distance between the triad’s global position (combined analysis) and each individual tissue in which the triad was considered expressed. We included only triads which were considered expressed in at least six tissues based on the combined analysis criteria outlined above. The average of these distances was defined as the “triad mean distance”. We ranked triads by their triad mean distance and the percentile was calculated by

$$\text{percentile}_i = \text{truncate}\frac{\text{rank}(\text{cmd}_i)}{\text{length}(\text{CMD})}$$

where CMD is the vector containing all the triad mean distance. The first and last deciles were classified as stable 10% and dynamic 10% triads, respectively. A similar approach was used to define the corresponding 5% and 25% extremes of the distribution. This analysis was conducted independently for the Chinese Spring no stress samples, the Azhurnaya developmental time course, and for each of the four tissue-specific networks. A visual representation is provided in fig. S19.

TE presence in gene promoters

We extracted all TEs that were annotated to fall at least partly within 1.5 kb and 5 kb upstream of the canonical ATG start-codon for all genes. We then split the TEs into the relevant gene lists covering homoeolog expression bias variation (stable 10%, middle 80%, and dynamic 10%) and homoeolog expression bias (balanced, dominant, nondominant, suppressed, and nonsuppressed) based on the “combined analysis (all tissues)” for CS. We used these lists to identify the proportion of genes and triads in each category which contained at least one TE in the promoter region.

Enrichment of TE families in gene promoters

Using the GFF file of TE coordinates, we extracted TEs present in the promoter regions of HC genes. We retrieved all TE copies that are entirely or partially present in the 5 kb upstream of the ATG start-codon of the canonical tran-script for each gene. We then calculated the number of genes in each of the stable 10%, mid-dle 80%, and dynamic 10% categories which contained specific TE families. We required the TE family to be present in at least 2% of the categorized genes for further analysis. We then found the deviation of this distribution from the expected 10-80-10 ratio using the c^2 test, P values adjusted with Benjamini-Hochberg. We calcu-lated the median length of each TE family based on all instances of that TE across the genome. We found fifteen TE families deviated signif-icantly from the expected 10-80-10 distribu-tion (Benjamini-Hochberg P < 0.01). However, the majority of these TE families were present in less than 5% of the genes considered, and showed very small variation in the number of promoters containing the TE, suggesting that the statistical significance may not be biologi-cally relevant.

TE density in gene promoters

We calculated the density of TEs within 5 kb upstream of genes by calculating the propor-tion of TE bases in sliding windows of 100 bp with a step size of 10 bp. The mean of each window was then calculated for both the stable 10%, middle 80%, and dynamic 10% triads and

the subgenome dominance categories (balanced, dominant, nondominant, suppressed, and non-suppressed). Mann Whitney tests with Benjamini-Hochberg adjusted P values were used to test for differences in TE density between catego-ries across each window.

Transcription factor binding site identification

The 1.5 kb of sequence upstream of the canonical ATG start-codon was used to identify transcription factor binding sites (TFBS) present in promoters of HC triads. The FIMO tool from the MEME suite [v 4.11.4 (63)] was used with a position weight matrix (PWM) obtained from plantPAN 2.0 (64) to predict TFBS based on previously identified sites across multiple plant species. FIMO was run with a P value threshold of <1E-04 (default), –motif-pseudo set to 1E-08 as recommended for use with PWMs and a –max-stored-scores of 1,000,000 to account for the large size of the dataset. The background model was generated from all extracted promoter se-quences using the MEME fasta-get-markov com-mand. Details of the TFBS comparisons between homoeologs is presented in the supplementary materials (13).

WGCNA network construction

Coexpression networks were built for six sepa-rate sample sets: grain, leaf, spike, root, abiotic and disease (table S1) using the WGCNA R pack-age (65). For each network, we selected HC genes which were expressed > 0.5 TPM in three or more samples. The count expression level of each gene was normalized using variance stabilizing trans-formation from DESeq2 (66) to eliminate differ-ences in sequencing depth between studies. The soft power threshold was calculated as the first power to exceed a scale-free topology fit index of 0.9 for each network separately. The soft powers used were: leaf = 12, spike = 12, roots = 7, disease = 7. For the abiotic and grain network the 0.9 threshold was not crossed until 15 and 20 re-spectively, which may be due to strong differ-ences between samples within these datasets, therefore the soft power threshold was selected according to the number of samples, resulting in abiotic = 7 and grain = 6. Signed hybrid net-works were constructed blockwise using the function blockwiseModules() with a maximum block size of 46,000 genes. The correlation type used was biweight mid-correlation "bicor" and the maxPOutliers was set to 0.05 to eliminate effects of outlier samples. The topographical overlap matrices (TOM) were calculated by the blockwiseModules function using TOMType = "unsigned" and the minimum module size was set to 30. The parameter mergeCutHeight = 0.15 was used to merge similar modules.

Defining same, similar, and divergent expression patterns of triads

For triads which had homoeologs within dif-ferent modules in the WGCNA networks we developed a threshold to determine whether the different modules had a "similar" or "divergent" expression pattern. We calculated the Euclidean distance between module eigengenes using the R package dist() and with these values we calculate the distances between the homoeologs in each triad. Triads where the pairwise distances were zero were in the same module. Triads where the pairwise distances were over zero were in different modules. For these triads in different modules when the pairwise distance between any two homoeologs was > 50% of the median maximum distance between eigengenes, the triad was classified as having a "divergent" expression pattern. In cases where the pairwise distance was over zero between at least one pair of homoeologs and the distance between all three pairs of homoeologs were =< 50% of the median maximum distance, the triad was classified as having a "similar" expression pat-tern. The median maximum distance between eigengenes was averaged across all four tissue networks to give a final threshold (50% of me-dian maximum distance) of 0.937431. This anal-ysis was carried out for 1:1:1 syntenic and 1:1:1 nonsyntenic triads expressed in each of the tis-sue networks (total triads = 9599 grain, 5378 leaf, 11,038 root, and 6173 spike). This excluded triads which had a putative transposable element (67 triads). Fig. S23 shows a graphic representation of this classification and the effect of altering the threshold in each of the four networks.

Module overlaps

Module overlap between networks was calcu-lated using the R package GeneOverlap which calculates significant overlaps between modules using a Fisher's exact test. Modules were con-sidered to have significant overlaps when the FDR adjusted P value < 0.05.

Correlation to stress status

Modules within the abiotic and disease networks were tested for correlations to intermediate level stresses using the cor(function). The significant of correlations were calculated using the func-tion corPvalueStudent() and corrected for multi-ple testing using p.adjust() using the Benjamini & Yekutieli method (67).

Genie3

HC genes expressed >5 TPM in at least one of the 850 sample were selected. Out of these 78,085 genes there were 3386 transcription factors (methods described above). Random forest regression was estimated for each gene based on the transcription factors as inputs using the genie3 package (68) in R (version 3.3.2) with default parameters (K=sqrt, nb.trees=1000, input.idx=list of transcription factors, importance. measure=IncNodePurity, seed=NULL). For each transcription factor, all predicted target genes (connectivity > 0.005) were extracted and func-tional enrichment within the target genes was determined using topGO (69) in R (version 3.3.2) with the following parameters (ontology = "BP", nodeSize = 10, classic Fisher test P < 10^{-10}). To summarize the results, the top three GO terms for each transcription factor, the P values for the strongest enrichment, and the direct blastx match in the Arabidopsis proteome (tair10) and well as the e-value and description were tabulated (table S28). The complete list of GO term enrichments of the biological process ontology for each tran-scription factor and the list of transcription factors associated with each GO term in the on-tology of biological process are published in e! DAL (https://doi.ipk-gatersleben.de/DOI/53148abd-26a1-4ede-802b-c2635af6a725/0dd8224a-34fc-4e3b-8ab8-883d07e52bd2/2/1847940088).

Identifying highly connected hub genes

Hub genes within each module for the abiotic and disease stress networks were calculated using the WGCNA R package function signedKME(). This calculates the correlation between the expression patterns of each gene and the module eigengene. Genes which were more highly correlated to the eigengene were considered hub genes.

REFERENCES AND NOTES

1. W. Albertin, P. Marullo, Polyploidy in fungi: Evolution after whole-genome duplication. Proc. Biol. Sci. 279, 2497–2509 (2012). doi: 10.1098/rspb.2012.0434; pmid: 22492065
2. S. P. Otto, J. Whitton, Polyploid incidence and evolution. Annu. Rev. Genet. 34, 401–437 (2000). doi: 10.1146/annurev. genet.34.1.401; pmid: 11092833
3. A. Salman-Minkov, N. Sabath, I. Mayrose, Whole-genome duplication as a key factor in crop domestication. Nat. Plants 2, 16115 (2016). doi: 10.1038/nplants.2016.115; pmid: 27479829
4. S. Renny-Byfield, J. F. Wendel, Doubling down on genomes: Polyploidy and crop plants. Am. J. Bot. 101, 1711–1725 (2014). doi: 10.3732/ajb.1400119; pmid: 25090999
5. M. Feldman, A. A. Levy, T. Fahima, A. Korol, Genomic asymmetry in allopolyploid plants: Wheat as a model. J. Exp. Bot. 63, 5045–5059 (2012). doi: 10.1093/jxb/ers192; pmid: 22859676
6. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. Nat. Rev. Genet. 18, 411–424 (2017). doi: 10.1038/nrg.2017.26; pmid: 28502977
7. S. Ohno, Evolution by Gene Duplication (Springer-Verlag, 1970).
8. F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. Biol. Sci. 279, 5048–5057 (2012). doi: 10.1098/rspb.2012.1108; pmid: 22977152
9. K. D. Makova, W.-H. Li, Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13, 1638–1645 (2003). doi: 10.1101/gr.1133803; pmid: 12840042
10. International Wheat Genome Sequencing Consortium, Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191 (2018). doi: 10.1126/science.aar7191
11. E. Martinez-Perez, P. Shaw, G. Moore, The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. Nature 411, 204–207 (2001). doi: 10.1038/35075597; pmid: 11346798
12. G. Moore, K. M. Devos, Z. Wang, M. D. Gale, Cereal genome evolution. Grasses, line up and form a circle. Curr. Biol. 5, 737–739 (1995). doi: 10.1016/S0960-9822(95)00148-5; pmid: 7583118
13. Additional materials and methods are available as supplementary materials.
14. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016). doi: 10.1038/nbt.3519; pmid: 27043002
15. P. Borrill, R. Ramirez-Gonzalez, C. Uauy, expVIP: A customisable RNA-seq data analysis and visualization platform. Plant Physiol. 170, 2172–2186 (2016). doi: 10.1104/ pp.15.01667; pmid: 26869702
16. M. Melé et al., The human transcriptome across tissues and individuals. Science 348, 660–665 (2015). doi: 10.1126/ science.aaa0355; pmid: 25954002
17. J. W. Walley et al., Integration of omic networks in a developmental atlas of maize. Science 353, 814–818 (2016). doi: 10.1126/science.aag1125; pmid: 27540173

18. M. Pfeifer et al., Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345, 1250091 (2014). doi: 10.1126/science.1250091; pmid: 25035498
19. D. Winter et al., An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLOS ONE 2, e718 (2007). doi: 10.1371/journal. pone.0000718; pmid: 17684564
20. P. Borrill, N. Adamski, C. Uauy, Genomics as the key to unlocking the polyploid potential of wheat. New Phytol. 208, 1008–1022 (2015). doi: 10.1111/nph.13533; pmid: 26108556
21. R. Avni et al., Functional characterization of GPC-1 genes in hexaploid wheat. Planta 239, 313–324 (2014). doi: 10.1007/ s00425-013-1977-y; pmid: 24170335
22. K. J. Simons et al., Molecular characterization of the major wheat domestication gene Q. Genetics 172, 547–555 (2006). doi: 10.1534/genetics.105.044727; pmid: 16172507
23. C. E. Grover et al., Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. 196, 966–971 (2012). doi: 10.1111/j.1469-8137.2012.04365.x; pmid: 23033870
24. S. Renny-Byfield et al., Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. Genome Biol. Evol. 6, 559–571 (2014). doi: 10.1093/gbe/ evu037; pmid: 24558256
25. M. Hao et al., The abundance of homoeologue transcripts is disrupted by hybridization and is partially restored by genome doubling in synthetic hexaploid wheat. BMC Genomics 18, 149 (2017). doi: 10.1186/s12864-017-3558-0; pmid: 28187716
26. M. J. Yoo, E. Szadkowski, J. F. Wendel, Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110, 171–180 (2013). doi: 10.1038/hdy.2012.94; pmid: 23169565
27. P. P. Edger et al., Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29, 2150–2167 (2017). doi: 10.1105/tpc.17.00010; pmid: 28814644
28. A. J. Bewick, R. J. Schmitz, Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36, 103–110 (2017). doi: 10.1016/ j.pbi.2016.12.007; pmid: 28258985
29. D. Zilberman, An evolutionary case for functional gene body methylation in plants and animals. Genome Biol. 18, 87 (2017). doi: 10.1186/s13059-017-1230-2; pmid: 28486944
30. A. J. Bewick et al., On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. U.S.A. 113, 9111–9116 (2016). doi: 10.1073/ pnas.1604666113; pmid: 27457936
31. X. Zhang et al., Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. PLOS Biol. 5, e129 (2007). doi: 10.1371/journal.pbio.0050129; pmid: 17439305
32. E. D. Akhunov et al., The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. 13, 753–763 (2003). doi: 10.1101/gr.808603; pmid: 12695326
33. R. J. A. Buggs et al., Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. New Phytol. 186, 175–183 (2010). doi: 10.1111/ j.1469-8137.2010.03205.x; pmid: 20409177
34. H. Zhao et al., Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. Plant Physiol. 176, 2789–2803 (2018). doi: 10.1104/ pp.17.01467; pmid: 29463772
35. C. D. Hirsch, N. M. Springer, Transposable element influences on gene expression in plants. BBA Gene Regul. Mech. 1860, 157–165 (2017).
36. A. Roulin et al., The fate of duplicated genes in a polyploid plant genome. Plant J. 73, 143–153 (2013). doi: 10.1111/ tpj.12026; pmid: 22974547
37. J.-T. Li et al., The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). Sci. Rep. 5, 8199 (2015). doi: 10.1038/srep08199; pmid: 25645996
38. C. Yu et al., MADS-box transcription factor OsMADS25 regulates root development through affection of nitrate accumulation in rice. PLOS ONE 10, e0135196 (2015). doi: 10.1371/journal.pone.0135196; pmid: 26258667
39. L.-H. Yu et al., MADS-box transcription factor AGL21 regulates lateral root development and responds to multiple external and physiological signals. Mol. Plant 7, 1653–1669 (2014). doi: 10.1093/mp/ssu088; pmid: 25122697
40. S. Guo et al., The interaction between OsMADS57 and OsTB1 modulates rice tillering via DWARF14. Nat. Commun. 4, 1566 (2013). doi: 10.1038/ncomms2542; pmid: 23463009
41. A. Ghazalpour et al., Integrating genetic and network analysis to characterize genes related to mouse weight. PLOS Genet. 2, e130 (2006). doi: 10.1371/journal.pgen.0020130; pmid: 16934000
42. M. Kumar et al., Heat shock factors HsfB1 and HsfB2b are involved in the regulation of Pdf1.2 expression and pathogen resistance in Arabidopsis. Mol. Plant 2, 152–165 (2009). doi: 10.1093/mp/ssn095; pmid: 19529832
43. K. M. Pajerowska-Mukhtar et al., The HSF-like transcription factor TBF1 is a major molecular switch for plant growth-to-defense transition. Curr. Biol. 22, 103–112 (2012). doi: 10.1016/ j.cub.2011.12.015; pmid: 22244999
44. M. Ikeda, N. Mitsuda, M. Ohme-Takagi, Arabidopsis HsfB1 and HsfB2b act as repressors of the expression of heat-inducible Hsfs but positively regulate the acquired thermotolerance. Plant Physiol. 157, 1243–1254 (2011). doi: 10.1104/ pp.111.179036; pmid: 21908690
45. G. Xu et al., uORF-mediated translation allows engineered plant disease resistance without fitness costs. Nature 545, 491–494 (2017). doi: 10.1038/nature22372; pmid: 28514448
46. K. V. Krasileva et al., Uncovering hidden variation in polyploid wheat. Proc. Natl. Acad. Sci. U.S.A. 114, E913–E921 (2017). doi: 10.1073/pnas.1619268114; pmid: 28096351
47. Y. Zhang et al., Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. Nat. Commun. 7, 12617 (2016). doi: 10.1038/ncomms12617; pmid: 27558837
48. FAO, www.fao.org/faostat/.
49. L. J. Leach et al., Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics 15, 276 (2014). doi: 10.1186/1471-2164-15-276; pmid: 24726045
50. N. Hamilton, ggtern: An extension to 'ggplot2', for the creation of ternary diagrams. R package version 2.2.1 (2016); https://CRAN.R-project.org/package=ggtern.
51. H.-B. Zhang, X. Zhao, X. Ding, A. H. Paterson, R. A. Wing, Preparation of megabase-size DNA from plant nuclei. Plant J. 7, 175–184 (1995). doi: 10.1046/j.1365-313X.1995.07010175.x
52. S. Lindgreen, AdapterRemoval: Easy cleaning of next-generation sequencing reads. BMC Res. Notes 5, 337 (2012). doi: 10.1186/1756-0500-5-337; pmid: 22748135
53. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572 (2011). doi: 10.1093/bioinformatics/btr167; pmid: 21493656
54. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2013); www.R-project.org/.
55. M. Dowle, M. Srinivasan, data.table: Extension of 'data.frame'. R package version 1.10.4-3 (2017); https://CRAN.R-project. org/package=data.table.
56. A. Akalin et al., methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 13, R87 (2012). doi: 10.1186/gb-2012-13-10-r87; pmid: 23034086
57. A. Akalin, V. Franke, K. Vlahoviček, C. E. Mason, D. Schübeler, Genomation: A toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics 31, 1127–1129 (2015). doi: 10.1093/bioinformatics/btu775; pmid: 25417204
58. H. Wickham, ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag, 2009).
59. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014). doi: 10.1093/bioinformatics/btu170; pmid: 24695404
60. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). doi: 10.1038/ nmeth.1923; pmid: 22388286
61. F. Ramírez et al., deepTools2: A next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975
62. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
63. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011). doi: 10.1093/bioinformatics/btr064; pmid: 21330290
64. C.-N. Chow et al., PlantPAN 2.0: An update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. Nucleic Acids Res. 44, D1154–D1160 (2016). doi: 10.1093/nar/gkv1035; pmid: 26476450
65. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). doi: 10.1186/1471-2105-9-559; pmid: 19114008
66. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). doi: 10.1186/s13059-014-0550-8; pmid: 25516281
67. Y. Benjamini, D. Yekutieli, The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann. Stat. 29, 1165–1188 (2001).
68. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods. PLOS ONE 5, e12776 (2010). doi: 10.1371/journal. pone.0012776; pmid: 20927193
69. A. Alexa, J. Rahnenfuhrer, topGO: Enrichment analysis for gene ontology. R package version 2.30.0 (2016). doi: 10.18129/B9. bioc.topGO

ACKNOWLEDGMENTS

We thank Bayer Crop Science staff members E. Caestecker and X. Wang for data analysis and B. Staelens, T. Debaecke, and A. Dobbelaere for plant growth and sampling. We also acknowledge the assistance of M. Burrell (NBI Computing) and A. Etuk (EI Digital Biology). Funding: This work was supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through the Designing Future Wheat (BB/P016855/1), GEN (BB/P013511/1), and Plant Health (BB/P012574/1) ISPs; Triticeae Genomics for Sustainable Agriculture (BB/J003557/1); ERA-PG (BB/G024960/1); ERA-CAPS (BB/N005007/1); and an Anniversary Future Leaders Fellowship to P.B. (BB/M014045/1). This work was also supported by the International Wheat Yield Partnership (IWYP76); the German Federal Ministry of Food and Agriculture (2819103915); the German Ministry of Education and Research (031A536); DFG (SFB924); Genome Canada/Ontario Genomics (OGI-128); the Canadian Applied Triticum Genomics project (CTAG2) funded by Genome Canada, Genome Prairie, Western Grains Research Foundation, Saskatchewan Wheat Development Commission, Alberta Wheat Development Commission, and Saskatchewan Ministry of Agriculture; the National Research Council of Canada Wheat Flagship program; the WGRCI/UCRC partially funded by NSF (IIP-1338897); French Agence Nationale de la Recherche grants ANR-11-BSV5-0015 and ANR-16-TERC-0026-01; and the European Research Council. S.A.H. was supported by the John Innes Foundation. C.J. was supported by Région Auvergne and the European Regional Development Fund (SRESRI 2016). This research was also supported in part by the NBI Computing infrastructure for Science (CiS) group through the HPC resources. The submission of sequencing data was brokered by the COPO platform (https://copo-project.org), funded by the BBSRC (BB/L024055/1), and supported by CyVerse UK, part of the Earlham Institute National Capability in e-Infrastructure. Author contributions: R.H.R.-G., P.B., and C.U. conceived, designed, and coordinated the study. R.H.R.-G. and P.B. organized RNA-seq samples and assigned metadata. R.H.R.-G. carried out the mapping and developed methods to analyze homoeolog expression patterns. R.H.R.-G. and C.U. analyzed homoeolog expression patterns and variation between tissues, cultivars, syntenic and nonsyntenic triads, chromosomal partitions, and progenitor species. P.B. constructed WGCNA coexpression networks, analyzed homoeolog coexpression, and identified biological case studies in the networks. P.B. carried out differential expression analysis between tissues. D.Lan., K.F.X.M., and M.S. generated gene annotations, performed phylogenomics analysis (including gene family, ortholog and homoeolog inference, phylogenetic trees, and TF superfamily classification), and analyzed subgenome expression bias. A.P. led informatics development, and Y.K. generated pictograph drawings for the wheat eFP portal, supervised by A.G.S. and N.J.P. A.Br. ran and analyzed the genie3 network. S.A.H. analyzed K_a/K_s ratios. A.T.C., S.J.R., and A.G.S. performed and analyzed DNA methylation profiles. D.Lat. performed ChIP-seq experiments, and L.C. performed bioinformatic analysis of ChIP-seq data, supervised by M.B. and A.Be. E.P. and C.J. analyzed histone marks. S.A.H., J.B., R.H.R.-G., and L.V. carried out the promoter analysis. T.F. illustrated wheat development for Fig. 1. E.P. identified chromosome partitions. F.C. identified and S.A.H. and J.B. analyzed transposable elements. J.B. conducted promoter cis-regulatory element analysis. M.D., J.J., F.v.E., S.J.R., A.T.C., H.S., B.S., D.X., C.J.R., B.C., B.B.H.W., R.A., V.T., R.D., C.J.P., and A.G.S. provided RNA-seq samples. P.B. and C.U. wrote the manuscript. R.H.R.-G., S.A.H., J.B., A.Br., D.Lan., A.T.C., L.C., C.J., E.P., and T.F. contributed text and figures for the manuscript. L.V., M.D., J.J., F.v.E., Y.K., J.B., H.S., B.S., C.J.R., R.A., V.T., R.D., F.C., C.J.P., N.J.P.,

A.G.S., and M.S. provided comments on the manuscript. All authors read and approved the final submission. Competing interests: The authors declare no conflicts of interest. Data and materials availability: All code used for the analyses of the datasets can be found at https://github.com/Uauy-Lab/ WheatHomoeologExpression, and data files are deposited in https://grassroots.tools. Sequencing reads were deposited with NCBI under accession codes PRJEB25639, PRJEB23056, PRJNA436817, SRP133837, PRJEB25640, and PRJEB25593 for RNA-seq; SRP133674 for whole-genome bisulfite sequencing; and SRP126222 for ChIP-seq. The full phylogenetic tree presented in Fig. 4C is available at http://itol.embl.de/shared/borrillp. Details of each dataset can be found in the methods and the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6403/eaar6089/suppl/DC1

Additional Materials and Methods
Figs. S1 to S24
Tables S1 to S37
IWGSC Collaborator List
References (70–97)
Data S1 and S2

30 November 2017; accepted 11 July 2018
10.1126/science.aar6089

CHAPITRE III

Paysage transcriptionnel chez le ble tendre : analyse des groupes homéologues en dyades et tetrades

Article II: Juery C., Concia L., De Oliveira R., Benhamed M., Ramirez-Gonzalez R., Uauy C., Choulet F. and Paux E. New insights into homoeologous copy number variations in the hexaploid wheat genome.

Submitted in Plant Genome.

Contexte

Ces travaux représentent la suite du premier article présenté dans le chapitre précédent. En effet, les travaux de Ramirez-Gonzalez *et al.* 2018 n'ont analysé les biais d'expression des gènes hoemoeologues que pour les triades, c'est-à-dire les 50% du génome. Ces gènes représentent la partie très conservée du génome, avec des fonctions plutôt représentatives du métabolisme de base, exprimés dans de très nombreuses conditions, conservés d'un point de vue synténique avec les gènes orthologues d'autres espèces. Or l'espace génique du blé tendre allohexaploïde est également caractérisé par un fort taux de gènes dupliqués mais comprend aussi 20% de ces gènes présentant une absence de l'une des copies sur l'un des sous-génomes (groupes d'homoeologie nommés « dyades »). Ainsi, nous avons souhaité explorer les biais d'expression et les caractéristiques de deux groupes de gènes faisant partie de ces deux catégories : les dyades (0 :1 :1 , 1 :0 :1, 1 :1 :0) et les tetrades comprenant une duplication sur chacun des sous-génomes (2 :1 :1, 1 :2 :2, 1 :1 :2). Ces groupes représentent deux forces évolutives majeures concernant l'évolution des gènes dans un génome polyploïde : la perte de gènes par fractionnement et la rétention de gènes dupliqués par néo/sous-fonctionnalisation.

Stratégie

Pour cela, nous avons utilisés les mêmes données que celles de l'article précédent à savoir les 123 échantillons de RNA-seq produits pour les 15 tissus en développement en conditions « normales » de culture et les données de ChIP-seq produites par l'IPS2 (équipe de Moussa Benhamed, publiées dans IWGSC 2018). Nous avons sélectionné les groupes de gènes dyades et tetrades sur la base des groupes définis par analyse phylogénétique, publiés dans IWGSC 2018. Pour l'analyse des biais d'expression, nous avons sélectionné les gènes présentant une moyenne d'expression, sur les 3 réplicas biologiques pour chacun des tissus, supérieure à 0,5TPM pour au moins un des 15 tissus.

Conclusions

Nous avons dans un premier temps retracé l'histoire évolutive des gènes en dyades et tétrades. Nous avons déterminé que très peu de gènes (environ 450 par sous-génomes, soit 1,3% de perte) absents pour les groupes des dyades correspondent à des gènes perdus post-polyploïdisation. Ils étaient donc déjà absents dans les génomes les espèces progénitrices (tetraploïde, diploïde génomes A et D). Idem pour les gènes dupliqués chez les tetrades, seuls 172 des 3008 tetrades sont des gènes dupliqués post-polyploïdisation. Ces éléments permettent d'interpréter les résultats concernant les données d'expression de façon plus éclairée.

Nous avons mis en évidence les dyades présentent pour 64% des groupes une expression équilibrée (« balanced »). Ainsi, ces gènes présentent des expressions conservées et certainement similaires chez les espèces progénitrices au même titre que les 81% de triades qui ne présentent aucun biais d'expression. Pour les 36% des groupes restant, le biais d'expression correspond à une suppression de l'une des copies, plus marqué pour les dyades AB. Ainsi, au cours des 800 000 d'évolution, certains de ces gènes ont pu acquérir une régulation différentielle (suppression de l'un des copies) permettant un retour à une expression proche de celle de l'espèce diploïde. Les tetrades quant à elles présentent une forte proportion de groupes de gènes présentant un biais d'expression (76% des groupes) correspondant en majorité à une suppression de l'un des paralogues. De plus, nous avons remarqué qu'une duplication sur le sous-génome A a plus d'impact sur la suppression de la copie B (et vice versa). Ainsi, une partie des biais d'expression des tetrades correspond à l'évolution des paralogues chez les espèces progénitrices (puisque ces gènes étainet majoritairement dupliqués avant la polyploïdisation) et une partie correspond à une évolution des sous-génomes A et B au sein de l'espèce tétraploïde durant 800 000ans (processus de néo/sous-fonctionnalisation).

Nous avons corrélé ces biais d'expression à un marquage épigénétique spécifique ainsi qu'au partitionnement des chromosomes : les dyades et les tétrades se retrouvent majoritairement au niveau des régions distales des chromosomes et présentent un marquage épigénétique plus fréquemment associé à la marque H3K27me3. Nous avons également remarqué que les gènes des sous-génomes A et B des tétrades et des dyades et les gènes des groupes présentant des biais d'expression sont plus fréquement associés à cette marque épigénétique. *A contrario,* les 81% des triades ne présentant pas de bias d'expression sont majoritairement couverts par la marque épigénétique H3K9ac, associée à une expression constitutive des gènes. H3K27me3, qui est associée à une expression transitoire et tissu-spécifique des gènes au cours du développement (hétérochromatine facultative) semble donc également jouer un rôle dans l'évolution des régulations transcriptionnelles post-polyploïdisation.

Pour conclure, nous proposons de distinguer les gènes homéologues du blé tendre en deux catégories : les gènes en trois copies, présentant une expression équilibrée et conservés au sein de l'évolution correspondraient aux gènes assurant les fonctions de base pour les cellules. Les gènes homéologues présentant une perte ou une duplication de copie seraient quant à eux des gènes évoluant plus rapidement, présentant des expressions et des foncitons spécifiques que nous avons qualifiés de gènes dispensables. A travers nos résutlats nous mettons également en évidence des biais d'expression plutôt liés à l'histoire évolutive des gènes chez les espèces progénitrices et intégration plus avancée des transcriptomes pour les sous génomes Aet B liée aux 800 000 ans de co-évolution au sein de l'espèce tétraploïde.

Implication personnelle

Ce travail constitue la partie du doctorat concernant les travaux réalisés en tant que premier auteur d'un article pouvant être publié dans une revue à comité de lecture. J'ai réalisé la plupart des analyses et ai écrit en entier le premier jet de l'article, par la suite remanié pour sa publication.

Received: 9 September 2020 | Accepted: 11 October 2020

DOI: 10.1002/tpg2.20069

The Plant Genome

ORIGINAL RESEARCH

New insights into homoeologous copy number variations in the hexaploid wheat genome

Caroline Juery[1] | **Lorenzo Concia**[2,4] | **Romain De Oliveira**[1] | **Nathan Papon**[1] | **Ricardo Ramírez-González**[3] | **Moussa Benhamed**[2] | **Cristobal Uauy**[3] | **Frédéric Choulet**[1] | **Etienne Paux**[1]

[1] Université Clermont Auvergne, INRAE, GDEC, Clermont-Ferrand 63000, France

[2] Institute of Plant Sciences Paris-Saclay (IPS2), UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, Orsay 91405, France

[3] John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

[4] Current address: Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, Université PSL, Paris 75005, France

Correspondence
Etienne Paux, Université Clermont Auvergne, INRAE, GDEC, 63000 Clermont-Ferrand, France.
Email: etienne.paux@inrae.fr

Funding information
European Regional Development Fund, Grant/Award Number: SRESRI 2016

Abstract

Bread wheat is an allohexaploid species originating from two successive and recent rounds of hybridization between three diploid species that were very similar in terms of chromosome number, genome size, TE content, gene content and synteny. As a result, it has long been considered that most of the genes were in three pairs of homoeologous copies. However, these so-called triads represent only one half of wheat genes, while the remaining half belong to homoeologous groups with various number of copies across subgenomes. In this study, we examined and compared the distribution, conservation, function, expression and epigenetic profiles of triads with homoeologous groups having undergone a deletion (dyads) or a duplication (tetrads) in one subgenome. We show that dyads and tetrads are mostly located in distal regions and have lower expression level and breadth than triads. Moreover, they are enriched in functions related to adaptation and more associated with the repressive H3K27me3 modification. Altogether, these results suggest that triads mainly correspond to housekeeping genes and are part of the core genome, while dyads and tetrads belong to the *Triticeae* dispensable genome. In addition, by comparing the different categories of dyads and tetrads, we hypothesize that, unlike most of the allopolyploid species, subgenome dominance and biased fractionation are absent in hexaploid wheat. Differences observed between the three subgenomes are more likely related to two successive and ongoing waves of post-polyploid diploidization, that had impacted A and B more significantly than D, as a result of the evolutionary history of hexaploid wheat.

1 | INTRODUCTION

Polyploidy, or whole genome duplication (WGD), has been observed in many living organisms, both prokaryotic and eukaryotic, and is widely recognized as a key driver of species evolution, diversification, as well as domestication

Abbreviations: GO, Gene ontology; HC, High confidence; HomoeoCNV, Homoeologous copy number variation; IWGSC, International Wheat Genome Sequencing Consortium; LC, Low confidence; PAV, Presence absence variation; PPD, Post-polyploid diploidization; TE, Transposable element; TPM, Transcripts per million; WGD, Whole genome duplication.

(Van de Peer, Mizrachi, & Marchal, 2017; Wendel, 2015). Polyploid species can be classified into two different categories: autopolyploids, which arise from genome doubling within one species, and allopolyploids, which arise from genome doubling following hybridization between two distinct species (Glover, Redestig, & Dessimoz, 2016). About one half of angiosperms are recent polyploids, including numerous important crop species such as oilseed rape (*Brassica napus*, 7500 years old), coffee (*Coffea arabica*, 10,000–50,000 years old) and wheat (*Triticum aestivum*, 10,000 years old). In addition, due to recurrent polyploidization events that occured through time, including the ζ (zeta) WGD event 300–350 million years ago (MYA), all flowering plants are ancient polyploids or paleopolyploids (Jiao et al., 2011). Well-characterized examples include sorghum (*Sorghum bicolor*, 95–115 MYA), maize (*Zea mays*, 26 MYA) and soybean (*Glycine max*, 13 MYA) (Qiao et al., 2019).

Each WGD event results in a doubling of the gene content. Salman-Minkov, Sabath, and Mayrose (2016) demonstrated that polyploidization gives fitness advantages through increasing the amount of raw genetic material on which natural and artificial selection can happen. However, despite the successive episodes of WGDs that occurred through time, the number of genes in plants is quite similar (Michael & Jackson, 2013) and far less than that expected by the doubling process (Adams & Wendel, 2005). This leads to the paradox that while being an important evolutionary process, polyploidy also seems to be an evolutionary 'dead-end' (Van de Peer et al., 2017) as polyploids systematically tend to return to a diploid state after a few million years (Wendel, 2015). Hence, duplicated genes can either be pseudogenized, silenced, and eventually lost or, alternatively, retained because having evolved a new function (neofunctionalization) or having diverged in expression (subfunctionalization) (Flagel & Wendel, 2009).

Previous studies on homoeologous gene loss and retention, as well as relative expression contribution in various polyploid species, revealed species-specific patterns, suggesting an effect of the age of the polyploidization and diploid progenitor divergence (Bottani, Zabet, Wendel, & Veitia, 2018). For example, in the paleopolyploid maize genome, 14% of coding sequences were lost during the diploidization process, with a 25% of differential loss between the two genomes and a biased fractionation (loss of functioning DNA sequence) in favor of one subgenome that exhibits an overall higher expression and higher impact on phenotypic variability (Jiao et al., 2017; Renny-Byfield, Rodgers-Melnick, & Ross-Ibarra, 2017). Similarly, in the ancient allotetraploid cotton genome, a biased fractionation was observed, with the A genome showing more gene loss, a faster evolution rate, and an overall lower expression level that the D genome (Zhang et al., 2015). In contrast, while frequent homoeolog sequence exchanges have been reported, no significant bias toward either subgenome was observed in the recent allotetraploid oilseed rape (*Brassica napus*) (Chalhoub et al., 2014). Therefore, for closely related progenitor genomes, like in soybean, a dosage sensitive pattern of expression leads to stochastic differentiation of homoeologous pairs. For highly divergent progenitor genomes, like maize, the more favorable homoeologous genes set of a subgenome are selected, leading to an overall subgenome retention and a biased fractionation.

Core Ideas

- Only one half of hexaploid wheat genes are in triads, i.e. in a 1:1:1 ratio across subgenomes
- Triads are likely part of the core genome; dyads and tetrads belong to the dispensable genome
- Subgenome dominance and biased fractionation are absent in hexaploid wheat
- Subgenome differences are related to two successive waves of post-polyploid diploidization

Bread wheat (*Triticum aestivum* L.) is an allohexaploid species (2n = 6X = AABBDD) originating from two successive rounds of hybridization (IWGSC, 2014; Marcussen et al., 2014). The first hybridization event occurred ~800,000 years ago between *Triticum urartu* (AA-genome) and an unknown *Aegilops* species (BB-genome). The second event took place ~10,000 years ago between *Triticum turgidum* (AABB-genome) and *Aegilops tauschii* (DD-genome). The resulting hexaploid AABBDD-genome was estimated to carry 107,891 high confidence (HC) protein-coding genes, although 161,537 low confidence (LC) genes and 303,818 pseudogenes and gene fragments were also annotated (IWGSC, 2018). Using a phylogenomics approach on a filtered set of 181,036 genes, 21,603 triads, defined as homoeologous genes that had a strict 1:1:1 correspondence (one copy per subgenome A, B and D), were identified. These account for only 36% of the gene set (64,809 genes) while the remaining 64% have a more complex homoeologous relationships (1:1:N or 0:1:1 for example). Similar proportions of genes in different homoeology contexts were observed on each of the subgenome. Together with equal contribution of the three homoeologous genomes to the overall gene expression, this supported the hypothesis of the absence of biased fractionation and global subgenome dominance (IWGSC, 2018). However, a cell type- and stage-dependent local subgenome dominance was observed (Harper et al., 2016; Pfeifer et al., 2014). A recent study reported that the vast majority of triads displayed a balanced contribution of each copy to the overall expression of the homoeologous group (Ramirez-Gonzalez et al., 2018). For those showing either dominance or suppression of one homoeologous copy, differences were associated with epigenetic changes,

especially in H3K9ac and H3K27me3 patterns. Such differences in gene expression likely represent the first steps toward neo- or subfunctionalization of wheat homoeologs.

Previous studies focused mainly on 1:1:1 triads leaving two third of the wheat genes apart. However, triads likely correspond to highly conserved and evolutionary constrained genes. In this regard, they may not be fully representative of the entire gene set and may not illustrate the complexity of the evolutionary trajectories that occurred within the hexaploid wheat genome. Here, we report on this unexplored part of the wheat genome by integrating not only triads but also dyads and tetrads, i.e. homoeologous groups that have undergone a single gene loss or duplication event, respectively. By combining genomic, transcriptomic and epigenetic data, we show that these two latter categories differ from triads not only by their chromosomal distribution but also by their transcriptional and epigenetic patterns as well as their conservation in wheat and other plant genomes, suggesting different evolutionary fates depending on the copy number of homoeologous genes.

2 | MATERIALS AND METHODS

2.1 | Definition and distribution of homoeologous group

Dyad, triad and tetrad gene information were retrieved from IWGSC (2018). Groups containing both high-confidence (HC) and low-confidence (LC) genes were filtered out to keep only those with HC genes. These data included gene position on the IWGSC RefSeq v1.0 reference sequence, homoeologous group category (dyad, triad or tetrad) and ID, as well as orthologous relationships with *Arabidopsis thaliana*, *Zea mays*, *Sorghum bicolor*, *Oryza sativa*, *Brachypodium distachyon* and *Hordeum vulgare* genes. The chromosome distribution of genes was performed by calculating the proportion of dyad, triad and tetrad genes over the total number of genes from this study within each of the five chromosomal regions defined by Pingault et al. (2015). Duplicated genes separated by less than 10 genes and less than 1 Mb on chromosomes were considered as tandem duplications. The other ones were considered as dispersed duplications.

2.2 | Characterization of ancestral duplications/deletions and presence-absence variations

To assess if deletions in dyads occurred within diploids or upon polyploidization, we aligned the two remaining copies onto diploid and tetraploid ancestor genomes using the GMAP package v2019.03.04 (Wu & Watanabe, 2005) with 85% of sequence identity and 85% of sequence coverage as parameters. For AB-dyads, we mined the *Aegilops tauschii* genome (Luo et al., 2017) with A and B coding sequences. For AD-dyads, we mined the B-genome of *Triticum dicoccoides* (Avni et al., 2017) with A and D coding sequences. For BD dyads, we mined the *Triticum urartu* genome (Ling et al., 2018), as well as the A-genome of *Triticum dicoccoides* with B and D sequences. To take into account polymorphisms between individual sequences (divergence, presence/absence variations…), we corrected these numbers by dividing them by the number of genes that are still present in the hexaploid wheat genome and that were found in the ancestral genomes (*e.g.* D-copy from an AD dyad in the *Ae. tauschii* genome). To assess if duplications in tetrads occurred within diploids or upon polyploidization, we used GMAP to estimate the number of copies in the diploid and tetraploid ancestor genomes. Presence/absence variation (PAV) analysis was performed as described by De Oliveira et al. (2020). Briefly, sequencing reads from 16 wheat accessions (Montenegro et al., 2017) were mapped on the IWGSC RefSeq v1.0 using BWA-MEM v0.7.12 (Li & Durbin, 2010). Alignments were then filtered using samtools view (samtools view -F 2308 -q11; Li et al., 2009) and PCR duplicates were removed using samtools rmdup. Depth of coverage was assessed using bedtools coverage (v2.26; Quinlan & Hall, 2010). Genes were considered as putative PAVs when their coding sequence was covered over less than 10% of their length in at least two accessions.

2.3 | Gene ontology enrichment and functional analysis

Gene Ontology (GO) terms and functional annotation data were retrieved from IWGSC (2018). GO enrichment analysis was conducted using R package *topGO* (Alexan & Rahnenfuhrer, 2019).

2.4 | Gene expression and relative contribution analysis

Expression data from 15 samples representing five different organs (root, leaf, stem, spike and grain) at three developmental stages each in controlled non-stressed conditions were retrieved from Ramirez-Gonzalez et al. (2018). Genes with expression levels below 0.5 TPM were considered as non-expressed. Outlier identification was conducted in R (R Core Team, 2014) using the *boxplot* function of the *ggplot2* package (Villanueva & Chen, 2019). They were defined as genes outside 1.5 times the interquartile range (IQR) above the third quartile ($Q3 + 1.5 \times IQR$) and below the first quartile ($Q1 - 1.5 \times IQR$).

For relative contribution analyses, we used the calculation method described by Ramirez-Gonzalez et al. (2018).

Briefly, to standardize the relative expression of each homoeolog across a group, we normalized the absolute TPM for each gene within this group, as follows:

$$\text{Expression of A} - \text{copy in an AB dyad, expression}_A = \frac{\text{TPM (A)}}{\text{TPM (A)} + \text{TPM (B)}}$$

$$\text{Expression of A} - \text{copy in an ABD triad, expression}_A = \frac{\text{TPM (A)}}{\text{TPM (A)} + \text{TPM (B)} + \text{TPM(D)}}$$

$$\text{Expression of A} - \text{copy in an ABDD tetrad, expression}_A = \frac{\text{TPM (A)}}{\text{TPM (A)} + \text{TPM (B)} + \text{TPM } (\text{D}_1) + \text{TPM } (\text{D}_2)}$$

The normalized expression was calculated for the average across all expressed tissues as well as for each tissue individually. In order to assign theoretical expression bias categories to each group within triad, tetrad and dyad, we constructed theoretical matrix (Supplemental Table S1). We calculated the Euclidean distance with the *rdist* function from R from the observed normalized expression of each group to each of the ideal categories. We assigned the homoeolog expression bias category for each group by selecting the shortest distance between theoretical and observed relative contribution values. For binary organ expression, genes expressed in an organ ($\geq$ 0.5 TPM) were given a value of 1 and those not expressed (<0.5 TPM), 0. This resulted in 32 binary expression profiles (0-0-0-0-0, 0-0-0-0-1, 0-0-0-1-1…).

2.5 | Histone mark analysis

Wheat H3K9ac and H3K27me3 data from bread wheat cultivar Chinese Spring at three-leaf stage were retrieved from IWGSC (2018). Genes were assigned a histone mark category, either H3K9ac, H3K27me3, both or no mark. We calculated meta-gene profiles for each category by computing the read density of each histone mark over different categories using Deeptools (Ramirez et al., 2016) computeMatrix scale-regions and plotted it with plotProfile. Only reads mapping within gene bodies plus 1 kb upstream of the transcription start site and 1 kb downstream of the transcription end site were considered. To account for different gene size, we divided the read counts over each gene by its length. H3K9ac and H3K27me3 data from *Oryza sativa* and *Zea mays* were retrieved from the Plant Chromatin State Database (Liu et al., 2018).

3 | RESULTS

3.1 | Defining homoeologous groups

To decipher the impact of gene loss and duplication in the wheat genome, we focused our study on high confidence (HC) genes from the IWGSC RefSeq v1.0. These 107,891 genes were previously assigned an homoeologous group defined through an iterative phylogenomic approach and for all of these groups, the cardinality was determined based on the number of homoeologs identified on each sub-genome (IWGSC, 2018). Using these data, we found 55,170 homoeologs (51.1%) belonging to 18,390 triads (Table 1; Supplemental Table S2; Supplemental Data S1). It is worth noting that an additional 2,218 triads containing both HC and LC genes were found. However, since LC genes corresponded to partially supported gene models, we did not include these triads in our analysis. We also found 12,640 genes (11.7%) corresponding to 6,320 groups having undergone a single gene loss during the course of evolution since the divergence of the A, B, and D subgenomes. The corresponding groups will be hereafter referred to as AB, AD and BD dyads, i.e. groups of HC genes being in 1:1:0, 1:0:1 or 0:1:1: ratios across homoeolog genomes, respectively. Finally, we identified 3,008 genes (2.8%) belonging to 240 AABD, 315 ABBD and 197 ABDD tetrads, i.e. groups having undergone one single gene duplication thus being in 2:1:1, 1:2:1 or 1:1:2 ratios, respectively. Similar to triads, 2,085 and 658 additional dyads and tetrads containing both HC and LC genes were found but not selected for further analyses. While in the present study we will focus only on dyads, triads and tetrads (70,818 genes), it is worth noting that 37,073 HC genes (34.4%) depart from these ratios, corresponding to genes that have undergone more than one deletion or duplication or that were not clustered into a homoeolog group. The overall high proportion of genes that are not in a strict 1:1:1 ratio, hereafter referred to as genes affected by homoeologous copy number variations (HomoeoCNVs), represent the dynamic part of the wheat genome during the course of its evolution either in the ancestral diploid species or after polyploidization.

3.2 | Conservation of homoeoCNVs

To assess whether dyad genes were lost upon or after polyploidization or were already missing in the progenitor genomes, we searched for orthologs in diploid and tetraploid genomes: *Triticum urartu* (AA-genome), *Triticum dicoccoides* (AABB-genome) and *Aegilops tauschii* (DD-genome). For the A-missing copies (i.e. BD dyads), we estimated that approximately 40.7% and 19.0% were still present in diploid and tetraploid ancestor genomes, respectively (Supplemental

TABLE 1 Number of groups and genes in the dyad, triad and tetrad categories

	Dyads			Triads	Tetrads			Total
	AB	AD	BD	ABD	AABD	ABBD	ABDD	–
Number of groups	1,776	2,253	2,291	18,390	240	315	297	25,462
	6,320			18,390	752			25,642
Number of genes	3,552	4,506	4,582	55,170	960	1,260	788	70,818
	12,640			55,170	3,008			70,818

Data S1). For the B-missing (i.e. AD dyads) and D-missing ones (i.e. AB dyads), the estimates were of 20.6% and 27.7% still present in tetraploid and DD-diploid ancestor genomes, respectively. These results revealed that most of the genes of the dyad category were already absent from the diploid and tetraploid progenitors and that roughly 450 genes were lost on each subgenome at each step of polyploidization (498 A genes from *T. urartu* to *T. dicoccoides*; 434 A genes from *T. dicoccoides* to *T. aestivum*; 465 B genes from *T. dicoccoides* to *T. aestivum*; 493 D genes from *Ae. tauschii* to *T. aestivum*). Similarly, for tetrads, the majority of genes duplicated in the hexaploid wheat genome were also found in two copies in the diploid or tetraploid genomes. Indeed, 65.8% and 76.3% of A-duplicates were found in two copies in *T. urartu* and *T. dicoccoides*, respectively, 67.3% of B-duplicates in *T. dicoccoides* and 83.2% of D-duplicates in *Ae. tauschii*. Overall, we estimated that 172 genes were duplicated upon hexaploidization, 29.1% on the A-genome, 52.9% on the B-genome and 18.0% on the D-genome. However, one cannot exclude that the absence of a gene is due to an intraspecific polymorphism.

To investigate the conservation of dyad, triad and tetrad genes in other bread wheat accessions, we mined for presence-absence variations (PAVs) of genes in the genome of 16 resequenced wheat accessions (Montenegro et al., 2017). Out of the 70,818 genes, we identified 2,270 putative PAVs representing 3.2% of the dataset (Supplemental Data S1). Consistent with the percentage of genes duplicated upon hexaploidization, the B-subgenome appeared to be more subject to variations (47.0% of all PAVs) than the A- and D-subgenomes (30.9% and 22.1%, respectively). When analyzing each category of homoeogroups individually, only 1.8% of triad genes were affected by PAVs whereas they accounted for 7.9% and 9.7% of dyads and tetrads, respectively. Interestingly, for tetrads, 74.1% of PAVs appeared to affect one of the duplicated homoeologs.

Finally, to look at the conservation of dyad, triad and tetrad genes in other plants, we used the orthologous relationships with *Arabidopsis thaliana*, *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, *Brachypodium distachyon* and *Hordeum vulgare* determined by the IWGSC (2018). The overall percentage of orthologs found for our 25,462 groups ranged from 52.8% in *A. thaliana* to 75.0% in *B. distachyon* (Supplemental Table S2). These proportions were consistent with the phylogenetic distance, the most distant species sharing the lowest number of orthologs, with the notable exception of barley, consistent with the a lower BUSCO score indicating the completeness of genome assembly, gene set and transcriptome calculated by the IWGSC (2018). When analyzing each category of homoeogroups separately, triad genes were found to be the most conserved (Figure 1). Indeed, the proportion of orthologous genes ranged from 58.4% in *A. thaliana* to 82.1% in *B. distachyon.* In tetrads, this proportion ranged from 37.6 to 56.4%. The least conserved genes were dyad ones with 32.0% of orthologs in *A. thaliana* and 48.3% in *B. distachyon.*

3.3 | Distribution of homoeoCNVs along wheat chromosomes

Previous studies revealed a partitioning of the wheat genome based on different structural and functional features, including the recombination rate, gene and transposable element (TE) densities, gene expression breadth, histone modifications, as well as gene and TE structural variation rate (Choulet et al., 2014; De Oliveira et al., 2020; IWGSC, 2018; Pingault et al., 2015). Consequently, chromosomes can be divided into five chromosomal compartments: the short arm distal R1, the short arm proximal R2a, the centromeric-pericentromeric C, the long arm proximal R2b and the long arm distal R3 regions. To investigate the distribution of HomoeoCNVs in the light of chromosome partitioning, we analyzed the proportions of each category (dyads, triads, and tetrads) in the proximal (R2 and C) and distal (R1 and R3) regions of the chromosomes (Figure 2a; Supplemental Table S2). We observed that triad homoeologs were more abundant in proximal than in distal regions: 64.3% *vs.* 35.7%, respectively. The opposite pattern was found for dyad genes with 62.2% located in distal regions and 37.8% in proximal regions. For tetrad genes, 57.6% were in distal and 42.4% in proximal regions.

At the chromosome scale, 95.6% of triads had their three genes located on homoeologous chromosomes (Figure 3). Out of the remaining 4.4%, 2.8% were found to have a mosaic distribution between chromosomes 4B, 4D and 5A or 4A, 5B and 5D, as a result of the structural evolution of the chromosomes 4A and 5A that have experienced inversions and translocations (Dvorak et al., 2018; Hernandez et al., 2012).

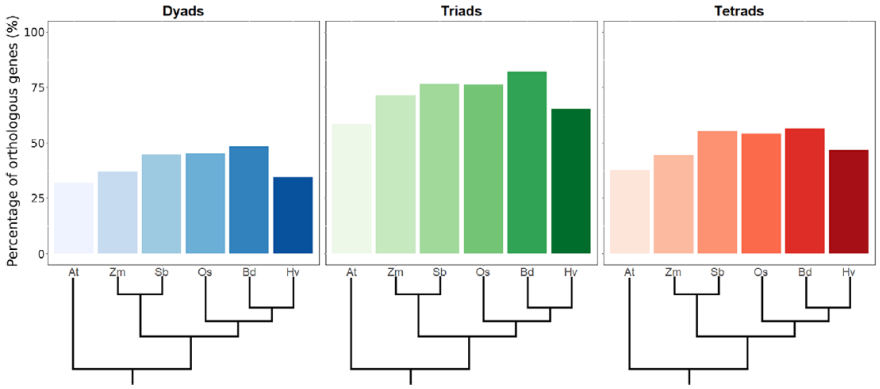


FIGURE 1 Conservation of genes in different plant genomes according to their category. For each category, the percentage of orthologs found in the *A. thaliana* (At), *Z. mays* (Zm), *S. bicolor* (Sb), *O. sativa* (Os), *B. distachyon* (Bd) and *H. vulgare* (Hv) genomes are given. Dyads are in blue, triads in green and tetrads in red

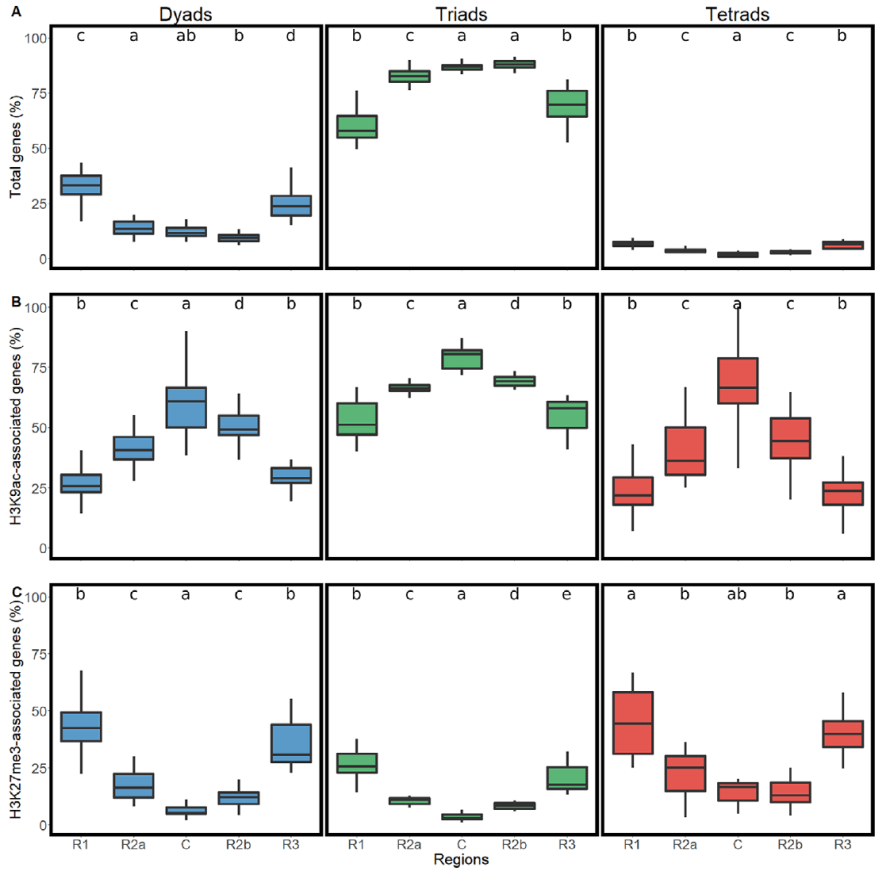


FIGURE 2 Distributions of genes of the different categories in the five regions of the wheat chromosomes. (a) Percentage of genes of a given category in a given region according to the total number of genes in this region. (b) Percentage of H3K9ac-associated genes according to the total number of genes from the same category in a given region. (B) Percentage of H3K27me3-associated genes according to the total number of genes from the same category in a given region. R1, R2a, C, R2b and R3 are the five chromosomal regions. Dyads are in blue, triads in green and tetrads in red. The boxplots depict the minimum without outliers, first quartile, median, third quartile and maximum without outliers. Different letters above the boxplots indicate significant differences ($P < .01$, Wilcoxon test)

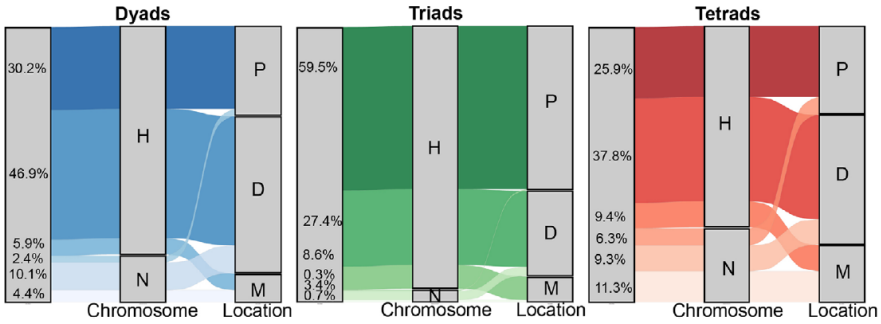


FIGURE 3 Alluvial plots of the different categories according to their location on chromosomes. For each category, the left-hand bar represents to whole set of genes, the central bar represents the position on homoeologous (H) or non-homoeologous (N) chromosomes; the right-hand bar represents the location, either proximal for all copies of a group (P), distal (D) or a mosaic of proximal and distal genes (M). The number indicated in the left-hand bar are the percentages of each class. Dyads are in blue, triads in green and tetrads in red

For dyads, 83.1% were found on homoeologous chromosomes and 3.3% showed a mosaic between chromosomes 4 and 5. For tetrads, the proportion of conserved homoeologous locations was much lower (73.1%) whereas that of mosaic distributions related to chromosome 4A evolutionary history was similar (3.2%).

As the boundaries of these regions are conserved between homoeologous chromosomes (IWGSC, 2018), we wondered to what extent the different gene copies of the same homoeologous group were located in the same regions (Figure 3). For triads, 90.7% of the genes belong to groups showing conserved locations for the three copies, with 30.8% being exclusively in distal regions and 59.9% being exclusively in proximal regions, confirming the high level of collinearity between A, B, and D. Only 9.4% of triad genes showed a variable location (mosaic distribution) between A, B, and D. For dyads, 57.0% of the homoeologs were exclusively located in distal regions, 32.7% were only in proximal and 10.3% were located in two different regions. For tetrads, 47.1% of the genes belong to groups having all their copies located exclusively in distal regions, 32.2% located exclusively in proximal regions and 20.7% with a mosaic distribution. The higher proportion of mosaic distribution for the tetrad category is explained by dispersed duplications that represented 36.6% of duplicated genes, of which 19.8% showed inter-chromosomal duplications.

3.4 | Functions of homoeoCNVs

Gene Ontology (GO) enrichment analysis revealed that dyads, triads and tetrads were involved in different biological processes. Indeed, triads were associated with basic cell processes such as transport, protein folding or DNA repair, replication and recombination. In contrast, tetrad and dyad genes were enriched in GO terms such as protein phosphorylation, oxidation-reduction processes, and response to fungus and oxidative stress (Supplemental Table S3). Analyzing the distal and proximal regions separately reached the same results, demonstrating that the GO enrichment was not only related to the preferential chromosomal location of the different categories (data not shown).

We expanded the analysis using the functional annotation of these genes to search for putative enrichment in protein functions (IWGSC, 2018) (Supplemental Data S1). F-box family proteins appeared to be the most abundant family in dyads and tetrads, comprising 7.9% and 6.0% of genes, respectively, while it represented 2.1% of triads. Similarly, consistent with GO enrichment analyses, disease resistance associated genes such as NLR, RLK, BTB/POZ-domain or ankyrin represented 9.7% of dyads and 7.8% of tetrads but only 2.9% of triads. Among other functions enriched in dyads and/or tetrads compared to triads were oxidation-reduction processes-associated proteins such as peroxidases, Cytochrome P450 and glutathione S-transferases.

3.5 | Expression of homoeoCNVs

To evaluate expression differences between the three categories of homoeologs, we used a gene expression atlas covering the whole plant development in controlled non-stressed conditions (Pingault et al., 2015; Ramirez-Gonzalez et al., 2018). We found detectable expression (TPM values >0.5) in at least one out of 15 tissues for 61,680 homoeologous genes from our dataset (87.1%) (Supplemental Table S2; Supplemental Data S1). The proportion of expressed genes was slightly higher on the D-genome genes (87.7%) than on the A- and B-genomes (86.6% and 86.9%, respectively; χ^2 p-value <.01). The percentage of expressed genes varied between categories too: 91.9% for triads (50,710 genes), 69.9% for dyads (8,838 genes) and 70.9% for tetrads (2,132 genes). In addition, we observed intra-category differences. For dyads, the AB-groups contained significantly

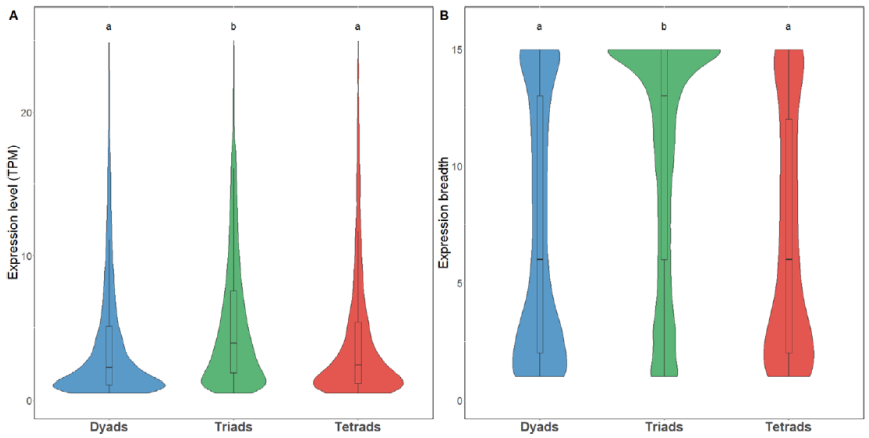


FIGURE 4 Expression level (a) and breadth (b) of the different categories. Expression level in TPM; expression breadth in number of conditions. Dyads in blue, triads, in green and tetrads in red. The boxplots depict the minimum without outliers, first quartile, median, third quartile and maximum without outliers. Different letters above the violin plots indicate significant differences ($P < .01$, Wilcoxon test)

fewer expressed genes (66.2%) than the BD- (70.5%) and AD-groups (72.2%) (χ^2 p-value <.01). For tetrads, at a χ^2 p-value of 1%, no significant differences were observed between groups. Interestingly, while in dyads and triads, the homoeologous genomes tended to have similar proportions of expressed copies, the duplicated-genome copies of tetrads displayed fewer expressed genes (χ^2 p-value <.01; Supplemental Table S2).

After discarding 7,179 outliers, i.e. genes with abnormally high expression level values (649 dyad, 6,375 triad and 155 tetrad genes), we investigated the mean expression level and expression breadth (i.e. the number of tissues in which genes were expressed) of 54,501 expressed genes: 8,189 dyad, 44,335 triad and 1,977 tetrad genes.

We found that triad genes were expressed at a higher level (mean = 5.9 TPM) and a higher breadth (10.4 tissues) than dyad (mean expression level = 5.2 TPM; mean expression breadth = 7.1) and tetrads (mean expression level = 5.2 TPM; mean expression breadth = 6.9) genes (Figure 4; Wilcoxon test p-value $<2.2\times10^{-16}$). To rule out the possibility that differences in expression level and breadth between dyads, triads and tetrads were only related to their chromosomal location, we divided the three categories in two sub-classes corresponding to their location, either proximal (R2 and C) or distal (R1 and R3). For all categories, genes located in proximal regions were expressed at significantly higher breadth than those in distal regions, confirming the impact of gene position on its expression. However, in general, triad genes were expressed at higher level and breadth than dyad and tetrad genes located in the same compartment (distal or proximal) (data not shown). This showed that the higher expression observed for triads may not only be due to their chromosomal location but to other factors. In tetrads, as for the proportion of expressed genes, the duplicated genome copies were less expressed than the non-duplicated ones, with lower expression breadth (6.6 vs. 7.1) and level (4.7 vs. 5.6; Wilcoxon test p-value <.01).

3.6 | Relative contribution of each copy to the expression of the overall homoeologous group

To go further on expression analysis of our three categories of homoeologs, we calculated the relative contribution of each homoeolog to the overall group expression, for groups having at least one gene expressed. We then assigned each group to an expression bias category, as defined by Ramirez-Gonzalez et al. (2018): the balanced category with similar relative abundance of transcripts from each of the homoeologs, and the homoeolog-dominant or homoeolog-suppressed categories, classified based on the higher or lower abundance of transcripts from a given homoeolog with respect to those from the other(s) (Supplemental Tables S2 and S3).

For dyads, 64.0% of the groups were balanced, while 36.0% were dominant/suppressed. AB dyads appeared to be less frequently balanced than AD and BD dyads (60.1%, 66.6% and 64.2%, respectively; χ^2 p-value <.05). The expression breadth of balanced dyads was higher than that of suppressed/dominant ones (8.0 and 4.9, respectively) (Figure 5).

For triads, an even higher proportion of balanced groups was observed (81.2%), while suppressed and dominant groups represented 14.0% and 4.8%, respectively. No difference was observed in the proportion of groups presenting a single-homoeolog dominance toward one sub-genome. Nevertheless, we observed a D-homoeolog suppression significantly less frequent (3.4%) than either A- or B-homoeolog suppression (5.3% and 5.2%, respectively; χ^2 p-value $<2.2\times10^{-16}$). As observed by Ramirez-Gonzalez and collaborators

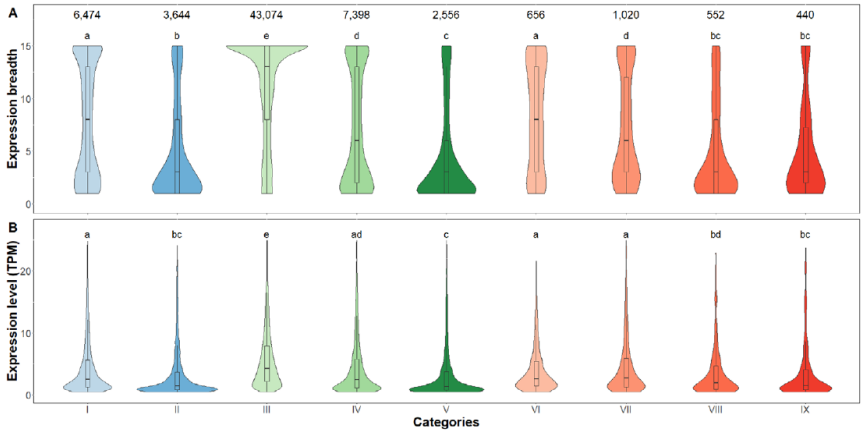


FIGURE 5 Expression breadth (a) and level (b) of the different categories according to their relative contribution classes. I: balanced dyads; II: Suppressed dyads; III: balanced triads, IV: Suppressed triads, V: Dominant triads, VI: balanced tetrads, VII: tetrads with one suppressed copy, VIII: tetrads with two suppressed copies, IX: tetrads with one dominant copy. The numbers above violin plots indicate the number of genes within the category. The boxplots depict the minimum without outliers, first quartile, median, third quartile and maximum without outliers. Different letters above the violin plots indicate significant differences ($P < .01$, Wilcoxon test)

(2018), expression breadth decreased from balanced to suppressed to dominant triads (11.1, 7.1 and 4.4, respectively) (Figure 5).

For tetrads, as expected from the greater number of gene copies, the pattern of relative contributions was much more complex. Balanced tetrads represented only 24.6% of the 667 groups having at least one expressed gene. The rest of the groups included 16.5% of groups with one copy dominant over the three others, 20.7% with two copies suppressed and 38.2% with one copy suppressed. It is worth noting that, for 74.5% of this latter, one of the two duplicates was suppressed. In addition, for ABDD tetrads, the duplication of a D-copy seems to have a similar impact on the suppression of A- and B-copies. By contrast, duplications of A-copies led to a slightly yet significantly higher proportion of B-copies than D-copies suppression (17.6% vs. 11.7%, respectively; χ^2 p-value <.01). A similar trend was observed in ABBD tetrads, where the B-copy duplication had greater impact on A-copy than D-copy suppression (15.7% *vs*. 11.2%, respectively; χ^2 p-value <.01). Interestingly, no significant difference was observed in terms of expression level between balanced tetrads and tetrads with one copy suppressed (5.2 TPM and 5.6 TPM, respectively) (Figure 5). The other tetrads displayed a significantly lower level (4.6 TPM for two suppressed copies and 4.6 TPM for one dominant copy; χ^2 p-value <.01).

We then explored whether the different categories retain their homoeologous expression bias category across the five organs (root, leaf, stem, spike and grain) (Supplemental Data S1). We found that 64.4% of balanced triads were also balanced (or not expressed) in the five organs, whereas, for dyads and tetrads, the proportions were 56.0% and 26.8%, respectively.

To complement this analysis, we investigated the divergence in spatial expression patterns. To this aim, we computed the binary expression (i.e. expressed or not) of each gene in the five different organs. This resulted in 32 binary expression clusters (0-0-0-0-0, 0-0-0-0-1, 0-0-0-1-1…). We then analyzed each group to see whether genes from a given group belong to the same or divergent binary expression groups (Supplemental Table S2).

For triads, 65.3% had their three copies in the same cluster, among which 85.8% were expressed in all five organs. When analyzing triads with one single divergent copy, we found a lower proportion of D-genome divergence, with 7.3% compared to 8.7% and 8.2% for the A and B-genomes, respectively (χ^2 p-value <.01).

For dyads and tetrads, the proportion of groups having all the genes in the same binary expression cluster dropped to 45.7% and 21.1%, respectively. Interestingly, 21.9% of tetrads had one single divergent copy and in 71.2% of the cases, the divergent copy was one of the duplicates. The proportion of D-divergent ABDD tetrads was found to be lower (63.9%) even though the difference was not significant, probably due to the small sample size.

Finally, when considering only balanced groups, the percentage of groups having all their copies in the same binary expression cluster raised to 64.8% for dyads, 75.4% for triads and 46.3% for tetrads.

3.7 | Epigenetic status of homoeoCNVs

Epigenetic marks, and especially H3K9ac and H3K27me3 histone modifications, have been shown to be associated

with differences in homoeolog expression patterns in triads (Ramirez-Gonzalez et al., 2018). These two marks have antagonist effects: H3K9ac is associated with open euchromatin and transcriptional activation whereas H3K27me3 is associated with facultative heterochromatin and transient transcriptional repression. To assess whether these marks may also be involved in the differences of expression patterns of dyads and tetrads, we analyzed the presence of these two marks on the 70,818 genes from our dataset. We found 44,954 genes associated with H3K9ac and 15,357 with H3K27me3 (Supplemental Data S1). After removing 3,809 genes that were associated with both marks, our dataset comprised 41,145 H3K9ac- and 11,548 H3K27me3-marked genes, i.e. 58.1% and 16.3% of all genes in the dataset, respectively (Supplemental Table S2). These proportions differed according to the chromosomal locations: distal and proximal regions were enriched in H3K27me3 and H3K9ac genes, respectively, as shown previously (IWGSC, 2018) (Figures 2b and 2c).

As expected, H3K27me3 genes tended to be more often repressed than H3K9ac, with 31.5% and 3.5% of genes never expressed across the 15 tissues, respectively. We also found a higher expression breadth for H3K9ac genes (12.1 tissues) compared to H3K27me3 genes (2.8 tissues). When analyzing gene expression in leaves at three-leaf stage (corresponding to ChIP-seq data), 85.1% of the H3K27me3 genes did not display any detectable expression while 83.2% of H3K9ac genes did.

When considering the three categories separately, we found differences in the proportion of the two marks. H3K27me3 was associated with 29.1% and 31.9% of dyad and tetrad genes, respectively, but only with 12.5% of triad ones (Figure 2b and 2c). By contrast, 64.7% of triad genes were marked by H3K9ac *vs.* 35.6% for dyads and 30.8% for tetrads.

Similar to what was observed previously, these proportions varied according to chromosomal regions (Figures 2b and 2c). However, triads were always more associated with H3K9ac and less with H3K27me3 than dyads and tetrads in the same chromosomal compartment. We also observed differences among tetrads. Indeed, the proportion of H3K9ac-associated genes increased from AABD to ABBD to ABDD (26.3%, 31.7% and 34.8%), while the opposite pattern was observed for H3K27me3 (34.7% for AABD, 31.7% for ABBD and 28.9% for ABDD). In addition, a significantly lower proportion of H3K9ac-associated genes was observed for the genome carrying the duplicated copies compared to the two others (27.7% vs. 33.9%; χ^2 p-value <.01). For H3K27me3, the proportion was slightly higher yet not significantly (33.1 vs. 30.8%; (χ^2 p-value >.01).

At the gene scale, balanced dyad, triad and tetrad genes had generally high H3K9ac and low H3K27me3 densities (Figure 6). For suppressed/non-suppressed and dominant/non-dominant groups, suppressed and non-dominant copies displayed higher H3K27me3 and lower H3K9ac than non-suppressed and dominant ones. Interestingly, the higher the number of suppressed copies in a group (from one to two in triads and from one to three in tetrads), the higher the H3K27me3 density, not only in the gene body but also into the upstream and downstream regions. This is consistent with the tight association of this mark with inactive promoters (Zhang et al., 2007).

We then analyzed whether genes from a group containing at least one gene associated with a mark tended to share the same epigenetic mark. For triads, 64.4% of the groups comprised three genes sharing the same mark, either H3K9ac (88.5%) or H3K27me3 (11.5%). This percentage was lower (55.9%) for dyads (57.1% H3K9ac and 42.9% H3K27me3). For tetrads, only 28.1% of groups comprised four genes sharing the same mark. Nevertheless, this percentage raised to 52.8% when including groups with three copies sharing the same mark.

Finally, we investigated the conservation of histone marks in two other species, *Zea mays* and *Oryza sativa*, for which H3K27me3 and H3K9ac data on a young leaf stage were available in the Plant Chromatin Database (Liu et al., 2018). Out of 12,995 and 12,006 groups containing at least one wheat H3K9ac-marked gene that had an ortholog in rice and maize, 10,541 (81.1%) and 7,604 (63.3%) were also marked by H3K9ac in these two species, respectively (Supplemental Table S2). For H3K27me3, the conservation was lower with 47.8% and 56.4% of the groups containing orthologs also targeted by this mark in rice and maize, respectively.

Surprisingly, while strong differences were observed at the sequence orthology level between dyads, triads and tetrads, the conservation of histone marks was not so different between the three categories. For example, while triads were much more conserved with rice than dyads (76.4% *vs.* 45.1%, respectively), quite similar proportions of groups containing conserved histone-marked genes were observed (48.4% of triads and 46.7% of dyads for H3K27me3, 81.5% of triads and 79.1% of dyads for H3K9ac).

4 | DISCUSSION

Wheat is an allohexaploid species originating from two successive and recent rounds of hybridization between three diploid species that were very similar in terms of chromosome number, genome size, TE content, gene content and synteny (IWGSC, 2018; Wicker et al., 2018). As a result, and considering that wheat is an autogamous homozygous species, it has long been considered that most of the genes were in three homoeologous copies. This perception started to change with the advent of the first draft assembly of the wheat genome sequence (IWGSC, 2014). The reference sequence of the hexaploid wheat genome confirmed that a significant fraction of genes departed from this 1:1:1 ratio and

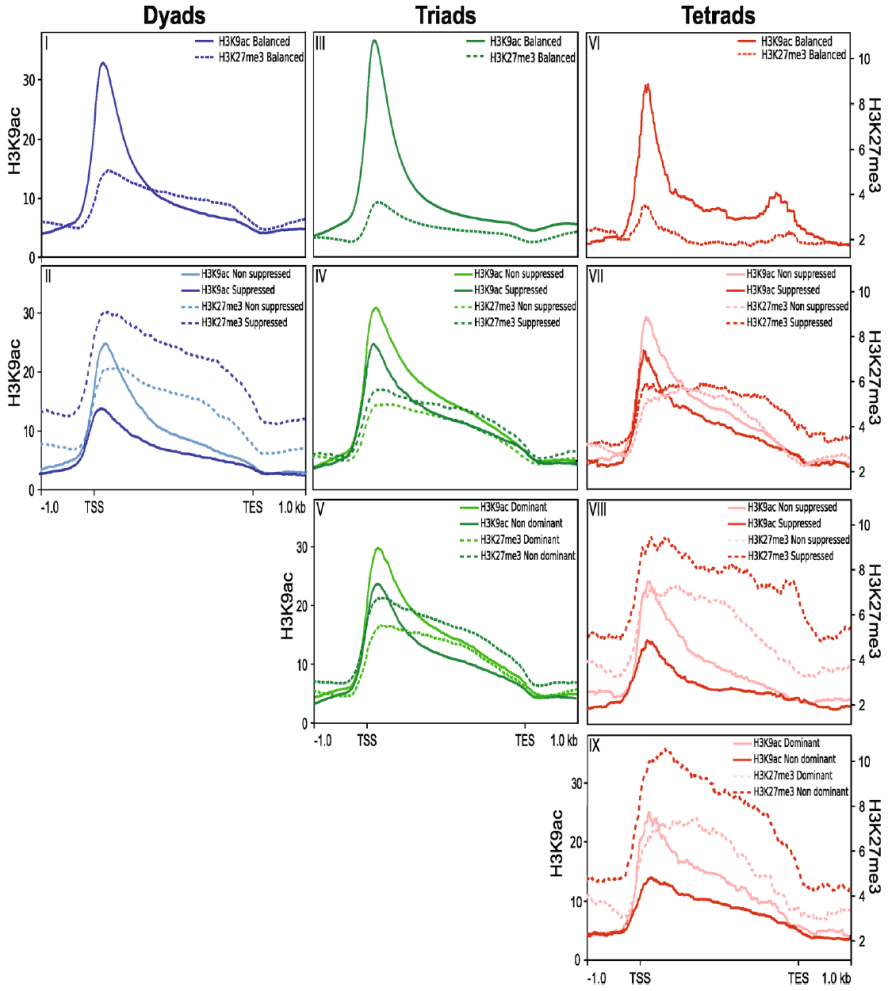


FIGURE 6 Epigenetic profiles of the different categories according to their relative contribution classes. Metagene profile for histone H3K9ac (solid lines) and H3K27me3 (dashed lines) marks from −1 kb upstream of the ATG to +1 kb downstream of the stop codon (normalized for gene length) for dyads (blue), triads (green) and tetrads (red) dominant/non-suppressed (light colors) and suppressed/non-dominant (dark colors) copies. I: balanced dyads; II: Suppressed dyads; III: balanced triads, IV: Suppressed triads, V: Dominant triads, VI: balanced tetrads, VII: tetrads with one suppressed copy, VIII: tetrads with two suppressed copies, IX: tetrads with one dominant copy

that these so-called 'triads' represent less than one half of all wheat genes.

In a recent work, Ramirez-Gonzalez et al. (2018) characterized the transcriptome atlas of wheat with a focus on these triads. This analysis provided new insights into the relative contribution of homoeologous copies to the overall group expression and the possible role of epigenetic marks in establishing this pattern.

In this study, we extended this analysis to genes departing from the 1:1:1 ratio, and more particularly the homoeologous groups having undergone a single gene loss or duplication event. These so-called dyads and tetrads, collectively referred to as HomoeoCNVs, represented 17.8% and 4.2% of our HC gene datasets whereas triads represented 77.9%. These proportions differed from those reported by the IWGSC (2018) as we focused our analysis on HC genes while the filtered dataset used by the IWGSC consisted of both HC and LC genes and took into account all categories, including genes in N:N:N ratio and not only dyads, triads and tetrads.

Because they have been kept in a strict 1:1:1 ratio through the course of evolution, triads are likely to correspond to highly conserved and evolutionary constrained genes. By contrast, dyads and tetrads have been either deleted or duplicated in the hexaploid wheat genome or in its diploid and tetraploid progenitors. We therefore suggest that triads are enriched in housekeeping genes and are part of the core genome, while dyads and tetrads belong to the dispensable genome of wheat. Several findings support this hypothesis.

First, we found that triads were more conserved in other plant genomes than HomoeoCNVs. By contrast, dyads and tetrads were found to be less conserved not only in distant plant genomes such as *A. thaliana*, *Z. mays*, *O. sativa*, *S. bicolor* or *B. distachyon* but also in the *Triticum/Aegilops* species, as most of these genes were already missing or duplicated in the wheat progenitors. In addition, HomoeoCNVs were also enriched in PAVs in a panel of 16 hexaploid wheat accessions. Previous studies in soybean, rice and *B. distachyon* demonstrated that core genes tend to have a higher percentage of homologs in other species than dispensable ones (Gordon et al., 2017; Li et al., 2014; Zhao et al., 2018). In wheat, this difference in terms of gene conservation is consistent with the genomic distribution of the different categories and the chromosome partitioning (Choulet et al., 2014; Daron et al., 2014; Darrier et al., 2017; Glover et al., 2015; IWGSC, 2018). Indeed, we showed that triads were more abundant in the low-recombination proximal regions. By contrast, dyads and tetrads were enriched in distal regions where differential TE content and recombination rate have likely driven gene duplications and deletions (Akhunov et al., 2003; Dvorak & Akhunov, 2005; Feldman, Levi, Fahima, & Korol, 2012; Reams & Roth, 2015; Zhang, 2003).

We also found that triads were expressed at higher level and breadth, while dyads and tetrads tend to be more specific to some tissues or developmental stages. In *B. distachyon*, core genes tend to be expressed at a higher level and more broadly than dispensable genes (Gordon et al., 2017). Choulet et al. (2014) and Pingault et al. (2015) reported on the physical partitioning of wheat genes, with highly and constitutively expressed genes being mainly located in proximal regions and genes expressed at lower level and breadth in distal ones. However, by analyzing distal and proximal regions separately, we showed that triads were always more expressed than dyads and tetrads whatever their position on the chromosome, which ruled out the possibility that the differences in expression patterns were only related to the chromosomal positions. Conversely, this difference can at least partly be explained by the epigenetic pattern of the categories of homoeologous genes. Indeed, triads were enriched in H3K9ac active euchromatin mark whereas dyads and tetrads were enriched in H3K27me3, a repressive mark related to facultative heterochromatin (Wiles & Selker, 2017). This differential association with active or repressive histones marks have already been reported in other species, such as potato where CNV frequency increased in genes lacking histone marks associated with permissive transcription (Hardigan et al., 2016). In wheat, we showed recently that genes affected by intra- and interspecific copy number variations were enriched in H3K27m3 (De Oliveira et al., 2020).

Finally, dyads and tetrads were enriched in functions associated with environmental and defence responses, a common feature of most plant dispensable genomes (Golicz et al., 2016; Gordon et al., 2017; Hurgobin et al., 2018; Li et al., 2014; McHale et al., 2012; Schatz et al., 2014). In particular, we found a higher proportion of genes associated to oxidation-reduction process that are known to be related to reactive oxygen species and putatively to biotic (pathogens) and abiotic (heavy metals, salt…) stress response mechanisms (Gullner, Komives, Király, & Schröder, 2018; Mir et al., 2015; Mittler, Vanderauwera, Gollery, & Van Breusegem, 2004; Veith & Moorthy, 2018). Disease resistance-associated families, such as NLRs, RLKs, ankyrin repeat or BTB/POZ domain-containing proteins were also found in higher proportions in dyads and tetrads than in triads (Sun, Zhu, Balint-Kurti, & Wang, 2020; Wang, Zou, Li, Lin, & Tang, 2020; Ye et al., 2017; Zhang et al., 2019).

We then examined intra-categories differences to investigate the possible impact of polyploidization on both core and dispensable genomes in wheat. Indeed, polyploidization is usually followed by a post-polyploid diploidization (PPD) process that tends to revert the polyploid genome into a quasi-diploid one (Mandáková & Lysak, 2018). PPD is accompanied by several mechanisms including gene neo/subfunctionalization, activation of transposable elements, epigenetic reprogramming and genome fractionation. Genome fractionation is a long-term process involving the loss of redundant genes and/or noncoding regulatory elements (Cheng et al., 2018). While it has been observed in several species and seems to be a common mechanism, differences have been observed according to the type of whole genome duplication (Garsmeur et al., 2013). In allopolyploids or paleo-allopolyploids such as *Arabidopsis thaliana*, maize (*Zea mays*), Chinese cabbage (*Brassica rapa*) and *Brassica oleracea,* duplicated genes are lost preferentially from one parental genome (biased fractionation) and the subgenome having retained the highest number of genes is more expressed (genome dominance) (Liu et al., 2014; Schnable, Springer, & Freeling, 2011; Wang et al., 2011). By contrast, in autopolyploids or paleo-autopolyploids, such as poplar (*Populus trichocarpa*) and pear (*Pyrus bretschneideri*), subgenome dominance is absent and genes tend to be evenly lost between the two subgenomes (Li et al., 2019; Liu et al., 2017). In wheat, some rapid changes following polyploidization have been reported, including chromosomal rearrangements, epigenetic changes or TE-related shift in centromere position (Badaeva, Dedkova, Pukhalskyi, & Zelenin, 2015; Dvorak et al., 2018; Jiao et al., 2018; Li et al., 2013; Liu et al., 2009; Shaked, Kashkush, Ozkan, Feldman, & Levy, 2001; Zhao et al., 2019). However, whether the wheat genome experiences subgenome dominance or biased fractionation is still a matter of debate. Different analyses reached contradictory results (El Baidouri et al., 2017; IWGSC, 2018; Pont & Salse, 2017).

Consistent with what was observed on a filtered set of 181,036 genes comprising both HC and LC genes (IWGSC, 2018), the number of genes analyzed in our study was highly

similar between subgenomes, with 23,411 on A, 23,524 on B and 23,883 on D. These similar proportions can be partly explained by the fact that the vast majority of these groups (75.9%) corresponded to triads, with one copy on each of the subgenomes. Such a high percentage of genes that are still present on the A, B and D-genomes demonstrate that no massive gene loss occurred upon polyploidization. Homoeologous groups that have lost one copy, i.e. dyads, represented 17.8% of our dataset. This category might reflect post-polyploidization gene loss. Interestingly, a lower number of AB-dyads (1,776) was observed compared to AD- or BD-ones (2,253 and 2,291, respectively). However, the analysis of the diploid and tetraploid ancestors suggested that only approximately 450 genes were lost on each subgenome at each step of polyploidization. This similar number of lost genes reveals the absence of a biased fractionation in wheat. Nevertheless, while no bias was found in terms of gene loss, we noticed subtle differences between genomes at the transcription and epigenetic levels.

The majority of triads displayed a balanced contribution of each copy to the overall group expression (81.2%). They also showed a high proportion of homoeologous genes having the same binary spatial expression (65.3%) and sharing the same histone mark (69.3%). However, we observed a slightly higher proportion of D-genome expressed genes compared to A and B, together with a lower proportion of D-suppressed triads, and a lower proportion of D divergent copies. This suggests a lower repression or subfunctionalization of D-genome homoeologs compared to their A and B counterparts. While these are very faint differences, they might explain the subtly yet significantly higher relative abundance of the D-subgenome transcripts (33.7%) compared to the A (33.1%) and B (33.3%) reported by Ramirez-Gonzalez et al. (2018).

Similarly, little differences were observed in dyads. They were mainly balanced (64.0%), with most of the homoeologs associated with the same epigenetic mark (65.1%). In addition, we found no bias in terms of dominance of one genome over the other. However, AB-dyads tended to be slightly less numerous and less expressed than the AD- and BD-ones, these two latter being similar on several aspects.

The pattern was much more complex for tetrads, with fewer balanced groups (24.6%) that might be explained by a higher proportion of mosaic patterns of epigenetic marks (71.9%). The overall higher proportion of H3K27me3- and lower proportion of H3K9ac-associated genes, especially on the genome carrying the duplicated copy, were likely related to subfunctionalization of retained paralogs as suggested by Makarevitch et al. (2013) and Berke, Sanchez-Perez, and Snel (2012). According to their mean expression level and breadth, as well as the epigenetic pattern, ABDD tetrads were more comparable to AB-dyads than to other tetrads. Indeed, the A and B copies of both ABDD tetrads and AB dyads were more associated with H3K9ac and less with H3K27me3, while the opposite pattern was found for those of AABD and ABBD tetrads. In addition, an extra copy on the B-genome (ABBD tetrads) appeared to have a much stronger impact on the A-genome than on the D-genome, while an extra D-copy had a similar impact on A and B in ABDD tetrads.

Based on these different results, we propose that the D-subgenome homoeologous genes are less repressed than the two others, and conversely, that their presence, either as single or duplicated genes, had a limited impact on the A- and B-copies. Differences observed between subgenomes are likely related to the D-genome more recent hybridization with the AABB tetraploid genome progenitor. This resulted in two successive PPD waves, that had impacted A and B more significantly since they spent more time together. However, unlike most of the allopolyploid species, subgenome dominance and biased fractionation are absent in hexaploid wheat. Indeed, while originating from the hybridization of three distinct species, the diploid donor genomes were very similar in terms of gene and TE contents prior to polyploidization. Consequently, individual genes, rather than subgenomes, experienced stochastic differences over longer periods of time, resulting in retention of the majority of WGD duplicates. In this regard, while being an allohexaploid species, wheat somehow resembles more to an autopolyploid in terms of evolutionary fate, as already observed in other paleo-allopolyploids such as soybean (*Glycine max*) and cucurbits (*Cucurbita maxima* and *Cucurbita moschata*) (Sun et al., 2017; Zhao, Zhang, Lisch, & Ma, 2017).

ACKNOWLEDGMENTS

We acknowledge Hélène Rimbert and Philippa Borrill for their assistance. C.J. was supported by Région Auvergne and the European Regional Development Fund (SRESRI 2016).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Lorenzo Concia https://orcid.org/0000-0002-7401-7214
Romain De Oliveira https://orcid.org/0000-0003-0017-6308
Etienne Paux https://orcid.org/0000-0002-3094-7129

REFERENCES

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, *8*, 135–141. https://doi.org/10.1016/j.pbi.2005.01.001

Akhunov, E. D., Akhunova, A. R., Linkiewicz, A. M., Dubcovsky, J., Hummel, D., Lazo, G., … Dvorak, J. (2003). Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 10836–10841. https://doi.org/10.1073/pnas.1934431100

Alexan, A., & Rahnenfuhrer, J. (2019). topGO: Enrichment analysis for gene ontology.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., … Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, *357*, 93–97. https://doi.org/10.1126/science.aan0032

Badaeva, E. D., Dedkova, O. S., Pukhalskyi, V. A., & Zelenin, A. V. (2015). Chromosomal changes over the course of polyploid wheat evolution and domestication. In Y. Ogihara, S. Takumi, & H. Handa (Eds.), *Advances in wheat genetics: From genome to field* (pp. 83–89). Tokyo: Springer.

Berke, L., Sanchez-Perez, G. F., & Snel, B. (2012). Contribution of the epigenetic mark H3K27me3 to functional divergence after whole genome duplication in *Arabidopsis*. *Genome Biology*, *13*, R94. https://doi.org/10.1186/gb-2012-13-10-r94

Bottani, S., Zabet, N. R., Wendel, J. F., & Veitia, R. A. (2018). Gene expression dominance in allopolyploids: Hypotheses and models. *Trends in Plant Science*, *23*, 393–402. https://doi.org/10.1016/j.tplants.2018.01.002

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., … Wincker, P. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, *345*, 950–953. https://doi.org/10.1126/science.1253435

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, *4*, 258–268. https://doi.org/10.1038/s41477-018-0136-7

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., … Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science*, *345*, 1249721. https://doi.org/10.1126/science.1249721

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., … Choulet, F. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology*, *15*, 546. https://doi.org/10.1186/s13059-014-0546-4

Darrier, B., Rimbert, H., Balfourier, F., Pingault, L., Josselin, A.-A., Servin, B., … Sourdille, P. (2017). High-resolution mapping of crossover events in the hexaploid wheat genome suggests a universal recombination mechanism. *Genetics*, *206*, 1373–1388. https://doi.org/10.1534/genetics.116.196014

De Oliveira, R., Rimbert, H., Balfourier, F., Kitt, J., Dynomant, E., Vrána, J., … Choulet, F. (2020). Structural variations affecting genes and transposable elements of chromosome 3B in wheats. *Frontiers in Genetics*, *11*, 891. https://doi.org/10.3389/fgene.2020.00891

Dvorak, J., & 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. https://doi.org/10.1534/genetics.105.041632

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C. M., Luo, M.-C., Deal, K. R., … McGuire, P. E. (2018). Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Theoretical and Applied Genetics*, *131*, 2451–2462. https://doi.org/10.1007/s00122-018-3165-8

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., … Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytologist*, *213*, 1477–1486. https://doi.org/10.1111/nph.14113

Feldman, M., Levi, A., Fahima, T., & Korol, A. (2012). Genomic asymmetry in allopolyploid plants: Wheat as a model. *Journal of Experimental Botany*, *63*, 5045–5059. https://doi.org/10.1093/jxb/ers192

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, *183*, 557–564. https://doi.org/10.1111/j.1469-8137.2009.02923.x

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., & Freeling, M. (2013). Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution*, *31*, 448–454. https://doi.org/10.1093/molbev/mst230

Glover, N., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Feuillet, C., & Choulet, F. (2015). Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. *Genome Biology*, *16*, 188. https://doi.org/10.1186/s13059-015-0754-6

Glover, N. M., Redestig, H., & Dessimoz, C. (2016). Homoeologs: What are they and how do we infer them? *Trends in Plant Science*, *21*, 609–621. https://doi.org/10.1016/j.tplants.2016.02.005

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., … Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390.

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., … Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, *8*, 2184. https://doi.org/10.1038/s41467-017-02292-8

Gullner, G., Komives, T., Király, L., & Schröder, P. (2018). Glutathione S-transferase enzymes in plant-pathogen interactions. *Frontiers in plant science*, *9*, 1836–1836. https://doi.org/10.3389/fpls.2018.01836

Hardigan, M. A., Crisovan, E., Hamilton, J. P., Kim, J., Laimbeer, P., Leisner, C. P., … Buell, C. R. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *The Plant Cell*, *28*, 388–405.

Harper, A. L., Trick, M., He, Z., Clissold, L., Fellgett, A., Griffiths, S., & Bancroft, I. (2016). Genome distribution of differential homoeologue contributions to leaf gene expression in bread wheat. *Plant Biotechnology Journal*, *14*, 1207–1214. https://doi.org/10.1111/pbi.12486

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Galvez, S., Schaaf, S., … Mayer, K. F. (2012). Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant Journal*, *69*, 377–386. https://doi.org/10.1111/j.1365-313X.2011.04808.x

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., … Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, *16*, 1265–1274.

International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat genome. *Science*, *345*, 1251788.

International Wheat Genome Sequencing Consortium (IWGSC) (2018). Shifting the limits in wheat research and breeding through a fully annotated and anchored reference genome sequence. *Science*, *361*, eaar7191.

Jiao, W., Yuan, J., Jiang, S., Liu, Y., Wang, L., Liu, M., … Chen, Z. J. (2018). Asymmetrical changes of gene expression, small RNAs and chromatin in two resynthesized wheat allotetraploids. *The Plant Journal*, *93*, 828–842. https://doi.org/10.1111/tpj.13805

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*, 524–527. https://doi.org/10.1038/nature22971

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., … dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*, 97–100. https://doi.org/10.1038/nature09916

Li, B., Choulet, F., Heng, Y., Hao, W., Paux, E., Liu, Z., … Zhang, X. (2013). Wheat centromeric retrotransposons: The new ones take a major role in centromeric structure. *The Plant Journal*, *73*, 952–965. https://doi.org/10.1111/tpj.12086

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., & Durbin, R. (2010). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K., … Zhang, S. (2019). Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). *Horticulture Research*, *6*, 34. https://doi.org/10.1038/s41438-018-0110-6

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., … Qiu, L.-J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*, 1045–1052. https://doi.org/10.1038/nbt.2979

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., … Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature*, *557*, 424–428.

Liu, B., Xu, C., Zhao, N., Qi, B., Kimatu, J. N., Pang, J., & Han, F. (2009). Rapid genomic changes in polyploid wheat and related species: Implications for genome evolution and genetic improvement. *Journal of Genetics and Genomics*, *36*, 519–528. https://doi.org/10.1016/S1673-8527(08)60143-5

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., … Paterson, A. H. (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, *5*, 3930. https://doi.org/10.1038/ncomms4930

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., … Su, Z. (2018). PCSD: A plant chromatin state database. *Nucleic Acids Research*, *46*, D1157–D1167. https://doi.org/10.1093/nar/gkx919

Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N., … Wang, X. (2017). Two highly similar poplar paleo-subgenomes suggest an autotetraploid ancestor of Salicaceae plants. *Frontiers in Plant Science*, *8*, 571–571.

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., … Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*, *551*, 498.

Makarevitch, I., Eichten, S. R., Briskine, R., Waters, A. J., Danilevskaya, O. N., Meeley, R. B., … Springer, N. M. (2013). Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27. *Plant Cell*, *25*, 780–793. https://doi.org/10.1105/tpc.112.106427

Mandáková, T., & Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, *42*, 55–65. https://doi.org/10.1016/j.pbi.2018.03.001

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., … Olsen, O. A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*, 1250092. https://doi.org/10.1126/science.1250092

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., … Stupar, R. M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, *159*, 1295–1308. https://doi.org/10.1104/pp.112.194605

Michael, T. P., & Jackson, S. (2013). The first 50 plant genomes. *The Plant Genome*, *6*, 1–7. https://doi.org/10.3835/plantgenome2013.03.0001in

Mir, A. A., Park, S.-Y., Sadat, M. A., Kim, S., Choi, J., Jeon, J., & Lee, Y.-H. (2015). Systematic characterization of the peroxidase gene family provides new insights into fungal pathogenicity in *Magnaporthe oryzae*. *Scientific Reports*, *5*, 11831.

Mittler, R., Vanderauwera, S., Gollery, M., & Van Breusegem, F. (2004). Reactive oxygen gene network of plants. *Trends in Plant Science*, *9*, 490–498. https://doi.org/10.1016/j.tplants.2004.08.009

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., … Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*, 1007–1013. https://doi.org/10.1111/tpj.13515

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., … Olsen, O. A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, *345*, 1250091. https://doi.org/10.1126/science.1250091

Pingault, L., Choulet, F., Alberti, A., Glover, N., Wincker, P., Feuillet, C., & Paux, E. (2015). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biology*, *16*, 29. https://doi.org/10.1186/s13059-015-0601-9

Pont, C., & Salse, J. (2017). Wheat paleohistory created asymmetrical genomic evolution. *Current Opinion in Plant Biology*, *36*, 29–37. https://doi.org/10.1016/j.pbi.2017.01.001

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., … Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, *20*, 38. https://doi.org/10.1186/s13059-019-1650-2

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team (2014). R: A language and environment for statistical computing. Vienna, Austria: Retrieved from www.r-project.org.

Ramirez-Gonzalez, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., … Uauy, C. (2018). The transcriptional landscape of polyploid wheat. *Science*, *361*. https://doi.org/10.1126/science.aar6089

Ramirez, F., Ryan, D. P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., … Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*, W160–165. https://doi.org/10.1093/nar/gkw257

Reams, A. B., & Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harbor perspectives in biology*, *7*, a016592–a016592. https://doi.org/10.1101/cshperspect.a016592

Renny-Byfield, S., Rodgers-Melnick, E., & Ross-Ibarra, J. (2017). Gene fractionation and function in the ancient subgenomes of maize. *Molecular Biology and Evolution*, *34*, 1825–1832. https://doi.org/10.1093/molbev/msx121

Salman-Minkov, A., Sabath, N., & Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, *2*, 16115. https://doi.org/10.1038/nplants.2016.115

Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., … McCombie, W. R. (2014). Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology*, *15*, 506.

Schnable, J. C., Springer, N. M., & Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 4069–4074. https://doi.org/10.1073/pnas.1101368108

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., & Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell*, *13*, 1749–1759. https://doi.org/10.1105/TPC.010083

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., … Xu, Y. (2017). Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Molecular Plant*, *10*, 1293–1306. https://doi.org/10.1016/j.molp.2017.09.003

Sun, Y., Zhu, Y. X., Balint-Kurti, P. J., & Wang, G. F. (2020). Fine-Tuning Immunity: Players and Regulators for Plant NLRs. *Trends in Plant Science*, *25*, 695–713. https://doi.org/10.1016/j.tplants.2020.02.008

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*, 411–424. https://doi.org/10.1038/nrg.2017.26

Veith, A., & Moorthy, B. (2018). Role of cytochrome P450s in the generation and metabolism of reactive oxygen species. *Current opinion in toxicology*, *7*, 44–51. https://doi.org/10.1016/j.cotox.2017.10.003

Villanueva, R. A. M., & Chen, Z. J. (2019). ggplot2: Elegant graphics for data analysis, 2nd edition. *Measurement-Interdisciplinary Research and Perspectives*, *17*, 160–167. https://doi.org/10.1080/15366367.2019.1565254

Wang, H., Zou, S., Li, Y., Lin, F., & Tang, D. (2020). An ankyrin-repeat and WRKY-domain-containing immune receptor confers stripe rust resistance in wheat. *Nature Communications*, *11*, 1353. https://doi.org/10.1038/s41467-020-15139-6

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., … Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, *43*, 1035–1039.

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *American Journal of Botany*, *102*, 1753–1756. https://doi.org/10.3732/ajb.1500320

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., … Choulet, F. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology*, *19*, 103. https://doi.org/10.1186/s13059-018-1479-0

Wiles, E. T., & Selker, E. U. (2017). H3K27 methylation: A promiscuous repressive chromatin mark. *Current Opinion in Genetics & Development*, *43*, 31–37.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Ye, Y., Ding, Y., Jiang, Q., Wang, F., Sun, J., & Zhu, C. (2017). The role of receptor-like protein kinases (RLKs) in abiotic stress response in plants. *Plant Cell Report*, *36*, 235–242. https://doi.org/10.1007/s00299-016-2084-x

Zhang, C., Gao, H., Li, R., Han, D., Wang, L., Wu, J., … Zhang, S. (2019). GmBTB/POZ, a novel BTB/POZ domain-containing nuclear protein, positively regulates the response of soybean to *Phytophthora sojae* infection. *Molecular Plant Pathology*, *20*, 78–91. https://doi.org/10.1111/mpp.12741

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, *18*, 292–298.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., … Chen, Z. J. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnology*, *33*, 531–537. https://doi.org/10.1038/nbt.3207

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., & Jacobsen, S. E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biology*, *5*, e129.

Zhao, J., Hao, W., Tang, C., Yao, H., Li, B., Zheng, Q., … Zhang, X. (2019). Plasticity in Triticeae centromere DNA sequences: A wheat × tall wheatgrass (decaploid) model. *The Plant Journal*, *100*, 314–327. https://doi.org/10.1111/tpj.14444

Zhao, M., Zhang, B., Lisch, D., & Ma, J. (2017). Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *The Plant Cell*, *29*, 2974–2994. https://doi.org/10.1105/tpc.17.00595

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*, 278–284. https://doi.org/10.1038/s41588-018-0041-z

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Juery C, Concia L, De Oliveira R, et al. New insights into homoeologous copy number variations in the hexaploid wheat genome. *Plant Genome*. 2020;e20069. https://doi.org/10.1002/tpg2.20069

CHAPITRE IV

Recherche méthodologique : optimisation de la technique ChIP-seq sur plusieurs tissus chez le blé tendre

Contexte

Au début du doctorat (2017), très peu de données « Whole Genome » n'avaient été produites pour étudier la distribution des marques histones et des états chromatiniens chez le blé tendre car la séquence de référence pour aligner les lectures de séquençage issues d'expérience de ChIP-seq manquait. En parallèle de la préparation de cette séquence génomique de référence, une équipe (David Latrasse, Lorenzo Concia au sein de l'équipe ChromD dirigée par Moussa Benhamed, Instite of Plant Science, Paris Saclay) a réalisé la production de données ChIP-seq pour quatre marques histones H3K36me3, H3K9ac, H3K4me3 et H3K27me3 sur un tissu (feuilles de plantules de 10 jours, stade 3 feuilles) pour la même accession que celle dont le génome a été séquencé (*T. aestivum* cv Chinese spring). Le projet de thèse visait au départ à produire dans la foulée de ces premières données des données de RNA-seq et de ChIP-seq pour la marque H3K27me3 sur une collection de tissus prélevé sur une cinétique de développement de *T. aestivum* cv Chinese spring et couvrant plusieurs stades et tissus (feuilles, grains, épis, tiges à différents stades).

Stratégie

N'ayant pas de technicien spécialisé ChIP-seq au laboratoire, nous avons utilisé le protocole de David Latrasse. Les tissus de la cinétique de développement avaient été prélevés avant la formation d'une semaine avec David Latrasse au sein de son laboratoire. Après avoir tenté de produire directement et en une seule fois des résultats pour l'ensemble des tissus au cours de cette formation, nous avons dû internaliser le protocole et tenter de l'optimiser pour les différents tissus afin d'obtenir des résultats reproductibles et comparables.

Conclusions

Aucun résultat n'a été obtenu de l'expérimentation ChIP-seq sur les différents tissus au cours de ce doctorat. En effet, l'hétérogénéité des tissus et le manque d'expérience sur la technique ont conduit à l'abandon du projet. Cependant, beaucoup de questions ont été soulevées quant à la difficulté de réaliser un atlas ChIP-seq sur tissus hétérogènes et je propose ici une réflexion quant aux points de vigilance à prendre en compte pour mettre en œuvre une telle expérimentation.

Implication personnelle

J'ai réalisé l'ensemble des expérimentations ChIP-seq dont les résultats sont présentés dans ce manuscrit : optimisation du protocole pour différents tissus. Je ne présente pas les résultats de RNA-seq produits au début du doctorat sur 15 conditions et qui devaient être analysés en parallèle des données de ChIP-seq.

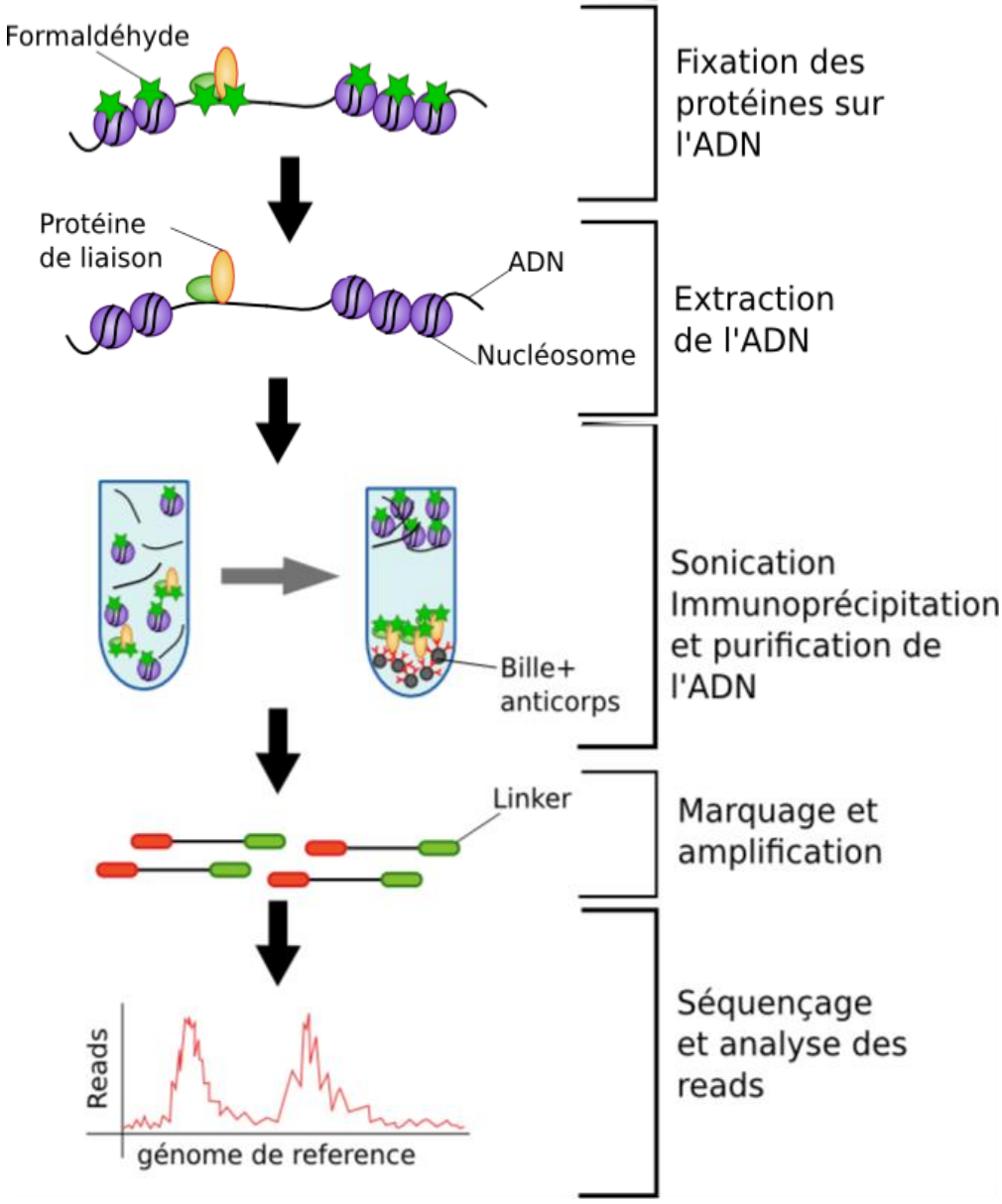


Figure 29. Schéma des étapes du protocole de ChIP-seq

I. Introduction

La technique de ChIP ou "Chromatin ImmunoPrecipitation" consiste en la capture d'une protéine adhérant à un locus (ou des loci) d'ADN sur loci génomiques grâce à un anticorps reconnaissant spécifiquement cette protéine (Raha *et al.* 2010). Après l'étape d'immunoprécipitation, la dissociation de la protéine des molécules d'ADN et la purification des fragments d'ADN permet ensuite de les séquencer. Le traitement bioinformatique par alignement de ces fragments sur la séquence génomique de référence de l'espèce étudiée permet d'obtenir la localisation génomique de la protéine étudiée. Les différentes étapes clefs du protocole de CHIP-seq sont présentées dans la figure 29. Les étapes cruciales de ce protocole qu'il est souvent nécessaire d'optimiser sont :

- La fixation (fixation des protéines à l'ADN dans les tissus, grâce à l'utilisation du formaldéhyde)
- La fragmentation de la chromatine pour obtenir de petits fragments correspondant à la taille d'un nucléosome permettant l'immunoprécipitation et le séquençage
- L'immunoprécipitation de la chromatine par la mise en présence de cette dernière avec un anticorps dirigé contre les marques épigénétiques d'intérêt

Jusque très récemment, les études à l'échelle génomique portant sur les marques histones chez les plantes ont été réalisées principalement sur *Arabidopsis thaliana* et les espèces de céréales modèles riz et maïs (Shi and Dawe, 2006; Saleh *et al.* 2008, Kaufmann *et al.* 2010, Gent *et al.*2012, Makarevitch *et al.*2013). Le site web http://systemsbiology.cau.edu.cn/ héberge une base de données répertoriant tous les résultats des études portant sur la chromatine de ces dernières années pour les plantes modèles Arabidopsis, Maïs et Riz, créée par l'équipe de Liu *et al*. en 2017. Si l'on regarde la diversité des tissus utilisés dans les expériences pour chaque espèce, on trouve 3 tissus et 15 « stades » de développement, 5 tissus et 11 « stades » et 7 tissus et 6 « stades » pour Arabidopsis, riz et le maïs respectivement. Or, très majoritairement ces études ne portent que sur 1 à deux tissus dans le même article. Citons l'exception de l'étude de Makarevitch *et al*. 2013 qui a étudié la distribution génomique de la marque H3K23me3 chez le maïs sur cinq tissus différents (épi mâle et femelles immatures, plantules de 12 jours et endosperme sur deux génotypes différents) mais en utilisant l'ancêtre du ChIP-seq, le ChOP-on-ChIP (hybridation des lectures de séquençage sur puce à ADN).

En 2017, première année du doctorat, une seule équipe avait produit des données ChIP-seq pour des marques histones chez le blé (H3K27me3, H3K36me3, H3K4me3 et H3K9ac) sur un seul tissu et sans réplica biologique : les plantules prélevées au stade 10 jours après germination (IWGSC 2018). Je citerai également l'article de Katie Baker sur l'orge (Baker *et al.* 2015) qui est l'espèce la plus proche du blé sur laquelle a été réalisée des expériences de ChIP-seq. Cette étude c'est concentré sur la caractérisation de plusieurs marques histones : H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me1, H3K27me2, H3K27me3, H3K36me3 and H3K56ac mais sur un seul tissu : feuilles de plantules au stade 10 jours après germination de la graine. Ainsi, à l'heure actuelle, aucune étude, dédiée spécifiquement à la réalisation

d'un atlas ChIP-seq sur une cinétique de développement n'a été réalisée chez les plantes, et encore moins chez une espèce polyploïde. Chez le blé, peu d'études ont été à l'heure actuelle publiées sur la caractérisation de la distribution des marques histones. Dans ce contexte, un des volets de ce travail de thèse a été d'internaliser un protocole de ChIP-seq et de tenter des optimisations pour produire des résultats de ChIP-seq pour la marque H3K27me3 sur plusieurs tissus chez le blé tendre. En parallèle de l'obtention de ces donnnées CHIP-seq, il était question de produire des données de RNA-seq pour corréler marquage épigénétique et expression des gènes.

II. Matériel

Les plantes utilisées pour la cinétique de développement correspondent à l'accession *Triticum aestivum cv Chinese spring* et cv Renan. Les 128 grains de la variété Chinese spring et les 8 grains de la variété Renan Renan nécessaires au plan d'échantillonnage (Tableau 3) ont été semés dans un terreau de semis utilisé en routine par les serristes du site de Crouël. 7 jours après le semis (plantules présentant des feuilles de 5 à 7 cm) les plantules ont été placées en vernalisation pendant 2 mois. Les conditions de température et d'éclairage durant cette période sont les suivantes : 17°C le jour (16h), 15°C la nuit (8h). Ensuite, les plantes ont été repiquées par deux dans des pots de 4 litres et placées dans une chambre de cultures en conditions contrôlées : éclairage jour/nuit correspondant à 16h/8h avec 21°C et 18°C de température. A partir de la sortie des anthères, les températures accumulées par chaque épi (comme des degrés jour) ont été notées afin de récolter les grains (50°J, 350°J et 500°J) aux mêmes stades de développement sur l'ensemble des plantes.

Stade	Tissus	Anticorps	Inputs
Levée_1	Feuilles	H3K23me3	Ok
Levée_2	Feuilles	H3K23me3	
3Feuilles_CS_1	Feuilles	H3K23me3	
3FeuillesCS_2	Feuilles	H3K23me3	
3Feuilles Renan_1	Feuilles	H3K23me3	Ok
3Feuilles Renan_2	Feuilles	H3K23me3	
2 nœuds_1	Tiges	H3K23me3	Ok
2 nœuds_2	Tiges	H3K23me3	
Floraison_1	Tiges	H3K23me3	
Floraison_2	Tiges	H3K23me3	
Meiose_1	Epi	H3K23me3	Ok
Meiose_2	Epi	H3K23me3	
Floraison_1	Epi	H3K23me3	
Floraison_2	Epi	H3K23me3	
50°J ap Anthèse_1	Grains	H3K23me3	Ok
50°J ap Anthèse_2	Grains	H3K23me3	
350°J ap Anthèse_1	Grains	H3K23me3	
350°J ap Anthèse_2	Grains	H3K23me3	
500°J ap Anthèse_1	Grains	H3K23me3	
500°J ap Anthèse_2	Grains	H3K23me3	
50°J ap Anthèse_1.1	Grains	H3K4me3	
50°J ap Anthèse_1.2	Grains	H3K4me3	
350°J ap Anthèse_1.3	Grains	H3K4me3	
350°J ap Anthèse_1.4	Grains	H3K4me3	
500°J ap Anthèse_1.5	Grains	H3K4me3	
500°J ap Anthèse_1.6	Grains	H3K4me3	
F1_Cs*Re 3Feuilles_1	Feuilles	H3K23me3	Ok
F1_Cs*Re 3Feuilles_2	Feuilles	H3K23me3	
F1_Re*Cs 3Feuilles_1	Feuilles	H3K23me3	Ok
F1_Re*Cs 3 Feuilles_2	Feuilles	H3K23me3	

Tableau 3. Plan de séquençage de données ChIP-seq sur plusieurs tissus

Les différents tissus (feuilles, tiges, épis, grains) ont été prélevés et plongés directement dans de l'azote liquide puis stockés à -80°C. Pour les grains, les épis ont été décortiqués sur de la glace et seuls les grains des épillets centraux ont été prélevés afin de respecter au maximum une cohérence de stade de développement des grains.

III. Méthode

Le protocole de ChIP-seq que nous avons utilisé au laboratoire a été celui utilisé par David Latrasse à l'IPS2 pour produire les premières données de ChIP-seq sur le blé tendre.

Je présente ici le protocole tel qu'il nous a été proposé et dans la description des résultats je décrirais les ajustements que nous avons testés. Les solutions du protocole à préparer sont présentées en annexe 1.

Crosslink du matériel végétal

1. Prélever 2 à 3g de matériel végétal et les placer dans un tube falcon de 50ml contenant 36ml H20.
2. Une fois que tous les échantillons sont prélevés, ajouter 1ml de formaldéhyde 37% dans chaque falcon, fermer les bouchons et mélanger (manipuler sous la hotte).
3. Placer à l'intérieur du tube des filtres nylons (ou languettes plastiques souples) permettant de garder immergé le matériel végétal pendant le crosslink sous vide.
4. Placer rapidement les échantillons sous vide pendant 15min.
5. Ajouter 2,5ml de Glycine 2M, mélanger, et replacer les tubes 5min sous-vide.
6. Vider la solution de formaldéhyde 1% dans une poubelle chimique adéquate (sans perdre le matériel végétal).
7. Rincer plusieurs fois les plantules avec de l'eau milliQ.
8. Après les rinçages, éliminer le plus d'eau possible du matériel végétal à l'aide de papier absorbant.
9. Congeler le matériel dans l'azote liquide (le matériel peut alors être conservé à -80°C).

Extraction de la chromatine

10. Préparer les tampons Extraction Buffer 1, Extraction Buffer 2, Extraction Buffer 3.
11. Broyer le matériel végétal en fine poudre dans l'azote liquide à l'aide d'un mortier et d'un pillon.
12. Pour 3g de matériel végétal, ajouter 25ml d'extraction Buffer 1 (environ 8ml/g) à la poudre dans un tube falcon de 50ml et homogénéiser la solution.
13. **Filtrer à l'aide de papier miracloth ou d'un tamis de filtration de 200 µm** dans un nouveau tube falcon. Répéter la filtration une seconde fois si nécessaire.
14. Centrifuger 15min à 4000rpm à 4°C.
15. Eliminer le surnageant.
16. Resuspendre le culot dans 25ml d'extraction Buffer 2. Incuber 10min sur glace.
17. Centrifuger 15min à 4000rpm à 4°C.
18. Eliminer le surnageant.
19. Resuspendre le culot dans 500µl d'extraction Buffer 3 (pipeter plusieurs fois pour reprendre le culot au mieux sans faire trop de mousse ; remarque : étape délicate ; il est possible d'utiliser un pinceau pour reprendre plus facilement le culot sans faire de mousse). Si l'échantillon est trop visqueux et ne se pipette pas correctement (dans le cas où le culot de noyaux et de débris était important), ajouter de l'Extraction Buffer 3 progressivement jusqu'à pouvoir pipeter l'échantillon correctement.
20. Dans 2 nouveaux tubes eppendorf de 1,5ml, ajouter 500µl d'extraction Buffer 3 (remarque : si le volume de l'étape 19 a été augmenté, prévoir 1 ou 2 tubes supplémentaires par échantillon – donc on a entre 2 à 4 tubes par échantillon suivant les quantités de culot obtenus / ces quantités varient en effet suivant le type de tissus
21. utilisés, et parfois on ne peut pas toujours peser exactement son matériel avant la manip.).

22. Prélever la solution contenant le culot resuspendu et les déposer doucement au-dessus des 2 tubes de 500µl d'extraction Buffer 3 de l'étape 20 (il faut répartir approximativement les échantillons resuspendus dans les 2 à 4 tubes préparés – on obtient une partie supérieure verte et une partie inférieure transparente).
23. Centrifuger 1h à 16000g à 4°C.
24. Préparer le Nuclei Lysis Buffer et le ChIP dilution Buffer.
25. Eliminer le surnageant.
26. Resuspendre chaque culot dans 300µl de Nuclei Lysis Buffer. La solution de chromatine obtenue devient visqueuse et elle doit pouvoir se pipeter correctement avec une pipette de 1ml. Si ce n'est pas le cas, il faut rajouter du Nuclei Lysis Buffer progressivement jusqu'à pouvoir pipeter correctement. (Si la chromatine est trop dense et trop visqueuse, la sonication ne se fera pas correctement). Pooler les 2 à 4 tubes pour chaque échantillon. La chromatine non soniquée peut alors être conservée à -20°C.

Sonication

26. Répartir 300 à 450µl de chromatine dans des tubes de 1,5ml pour la sonication. Soniquer la chromatine à l'aide du Bioruptor (Diagenode) 30sec ON / 30sec OFF puissance High pour un total de 60 cycles (60 minutes). (remarque : je sonique d'abord 2 fois 10min avec les tubes spéciaux pour le sonicateur qui sont plus rigides mais qui se détériorent, puis pendant 40min avec des tubes classiques).
27. Centrifuger 5min 13000rpm à 4°C.
28. Transférer le surnageant dans un nouveau tube. La chromatine soniquée peut ainsi être conservée à -20°C.
29. Prélever un aliquot pour la préparation ultérieure de l'Input (entre 1 et 10% du volume d'extraits).
30. Vérifier l'efficacité de la sonication avant de poursuivre :

- Préparer un gel d'agarose 1%.
- Prélever 25µl de chromatine soniquée et compléter le volume à 50µl avec le Nuclei Lysis Buffer.
- Ajouter 1µl de Cocktail RNase A+T1 (Ambion) et incuber 20min à 37°C.
- Ajouter 1µl de 0,5M EDTA, 2µl de Tris-HCl 1M (pH 6,5), et 2µl de protéinase K (20mg/ml Invitrogen) et incuber 1h à 65°C.
- Ajouter 50µl de phenol/chloroforme et vortexer 30sec.
- Transférer la phase aqueuse dans un nouveau tube.
- Ajouter 1µl de Cocktail RNase A+T1 (Ambion) et incuber 20min à 37°C.
- Déposer les échantillons sur gel (remarque : à la place du tampon de charge avec le bleu de bromophénol, j'utilise du glycérol 30% afin de ne pas avoir de problèmes pour observer le smear).
- Le smear doit être compris entre 150 et 500pb. S'il reste des fragments d'ADN supérieurs, resoniquer le surnageant sans les débris pendant 5 à 10min maximum, puis revérifier à nouveau l'efficacité de sonication sur gel.

Immunoprécipitation

31. Si la sonication est correcte, la chromatine soniquée peut alors être utilisée pour l'immunoprécipitation (IP). Pour chaque IP, déposer 100µl à 120µl de chromatine dans un tube eppendorf de 1,5ml et diluer 10 fois l'échantillon à l'aide du ChIP dilution Buffer.
32. Ajouter l'anticorps (entre 2 et 10µg en général par IP) et incuber O.N à 4°C en rotation.

33. Laver 50µl de Billes Protein A ou G (choix en fonction de l'anticorps utilisé) par IP, 3 fois avec 1ml de ChIP Dilution Buffer.
34. Ajouter les extraits contenant les immunocomplexes aux tubes contenant les billes.
35. Incuber 1h à 3h à 4°C en rotation.
36. Jeter le surnageant.

Lavages

37. Deux possibilités pour les lavages :

- Pour un ChIP « histones », laver les billes 2 fois avec 1ml de Low Salt Wash Buffer, puis 2 fois avec 1 ml High Salt Wash Buffer, 2 fois avec 1ml de LiCl Wash Buffer, puis 2 fois avec 1ml de TE Buffer.
- Pour un ChIP autre que « histones », laver les billes 6 fois avec 1 ml de ChIP Dilution Buffer, puis 2 fois avec 1ml de TE Buffer.

Elution, reverse crosslink, et purification de l'ADN

38. A partir de là, deux cas sont possibles :

- Dans le cas d'un ChIP-qPCR, l'élution, le reverse crosslink et la purification de l'ADN sont effectués à l'aide du kit « Ipure Kit » (Diagenode) en suivant exactement le protocole décrit dans le kit. L'ADN peut alors être utilisé pour les qPCR.

- Dans le cas d'un ChIP-seq :

39. Eluer les complexes en ajoutant 200µl de Elution Buffer. Vortexer et incuber 15min à 65°C en vortexant 2 à 3 fois pendant l'incubation.
40. Centrifuger brièvement et transférer le surnageant dans un nouveau tube.
41. Répéter l'élution avec à nouveau 200µl d'Elution Buffer.
42. Combiner les 2 éluats.
43. Sortir les INPUT et ajuster leur volume à 400µl avec Elution Buffer.
44. Ajouter 16µl de NaCl 5M aux échantillons IP et INPUT et incuber à 65°C O.N pour le reverse crosslink.
45. Ajouter 4µl de Cocktail RNase A+T1 (Ambion) et incuber 30min à 37°C.
46. Ajouter 8µl de 0,5M EDTA, 16µl de Tris-HCl 1M (pH 6,5), et 4µl de protéinase K (20mg/ml Invitrogen) et incuber 3h à 50°C.
47. Ajouter 450µl de phenol/chloroforme et vortexer 30sec.
48. Centrifuger 10min à 13000rpm.
49. Transférer 400µl de la phase aqueuse dans un nouveau tube.
50. Ajouter 40µl de Na acetate 3M pH5,5 et 1µl de Glycogen ou Glycoblue (Invitrogen 15mg/mg) et mélanger. Ajouter 1ml d'ethanol 100% et incuber à -20°C O.N.
51. Centrifuger 20min à vitesse max.
52. Jeter le surnageant et laver le culot avec 800µl d'EtOH 70%.
53. Laisser sécher le culot jusqu'à ce qu'il n'y ait plus de traces d'éthanol.
54. Reprendre le culot dans 10 à 20µl H20.

Quantification des ADN précipités et Input à l'aide du Qubit et du kit « Qubit dsDNA HS Assay Kit » (Invitrogen) selon le protocole du fabricant.
Synthèse de la banque pour un séquençage Illumina : La banque est préparée à partir de 10ng d'ADN à l'aide du kit « NEBNext Ultra II DNA library prep Kit» (NEB) et de 10 à 12 cycles PCR.

IV. Résultats

Dans les paragraphes sont présentés les résultats obtenus au cours de la phase d'optimisation du protocole de ChIP-seq, principalement pour les étapes de fixation au formaldéhyde et de fragmentation de la chromatine testés pour deux tissus contrastés : les feuilles et les grains. L'objectif de la présentation de ces résultats est de proposer une réflexion sur la difficulté de calibrer cette technique pour obtenir des données de ChIP-seq sur des tissus hétérogènes qui puissent être comparés.

IV.1 Prélèvement des tissus

En parallèle du développement de l'atlas de ChIP-seq il était prévu d'obtenir des données RNA-seq sur les mêmes tissus, issus des plantes de la même cinétique de développement pour pouvoir relier expression des gènes et statut épigénétique. C'est pourquoi, les échantillons pour les deux analyses ont été prélevés au même moment et ont été congelés à -80°C. Ayant commencé par réaliser les expérimentations RNA-seq sans avoir au préalable optimisé la technique de ChIP, et étant donné le grand nombre de tissus à prélever pour des stades de développement parfois très rapprochés, il était difficile de récolter les tissus frais pour réaliser directement les premières étapes de fixation-sonication-extraction du protocole de ChIP. Selon plusieurs protocoles et forums dédiés, il est conseillé que les premières étapes du ChIP (fixation, extraction, sonication) se fassent extemporanément à la récolte des tissus, donc sur des tissus frais. Or, n'ayant pas connaissance de cette limite lors du prélèvement des tissus pour le RNA-seq, les tissus prélevés pour le ChIP-seq ont aussi été congelés ; ils ont alors été fixés en les plongeant dans la solution de formaldéhyde 1% à température ambiante sous vide pendant 15 minutes. La fixation s'est donc réalisée en parallèle de la décongélation progressive des tissus. Nous n'avons pas testé la possible implication de l'effet congélation/décongélation sur la réussite de l'expérience.

IV.2 Tests pour l'optimisation des étapes de fixation et fragmentation

Dans un premier temps, nous avons réalisé les expériences de ChIP au sein du laboratoire de l'équipe ChromD à l'IPS2 avec David Latrasse en suivant son protocole. Nous avons obtenu de très faibles quantités d'ADN à la toute fin de l'expérience avec parfois des quantités très différentes entre les échantillons d'un même tissu (notamment les tissus stades 3 feuilles, tiges et épis) (annexe 2). Ceci indiquait une faible répétabilité du protocole en l'état, étant donné les caractéristiques contrastées des différents tissus, et probablement également due à la multiplication du nombre d'expérimentateurs (au nombre de 4 à ce moment-là). Nous avons par la suite, avec Jonhatan Kitt, technicien de l'équipe, essayé d'internaliser le protocole dans le laboratoire d'accueil de la thèse. Nous avons décidé d'utiliser le même principe de fragmentation de la chromatine que celui utilisé à l'IPS2, à savoir une sonication par des ondes électromagnétiques en utilisant un appareil de fragmentation de l'ADN : le covaris M220. Ne disposant pas du même appareil que l'équipe de Paris (Covaris S220) la première étape a été de tester les paramètres de sonication de notre appareil.

Pour mettre au point les temps de sonication pour les échantillons de la cinétique, nous avons d'abord travaillé sur d'autres échantillons, tels que des feuilles fraîches (stade pré-floraison) de la variété Renan prélevées sur des plantes cultivée en serre. Nous avons utilisé différents poids de matériel de départ, fixé 15 minutes sous vide avec du formaldéhyde 1% et stoppé la réaction de fixation avec de la Glycine 2M (cf protocole dans Methode). Ce premier test de fragmentation a été réalisé avec 10 cycles de sonication/pause (aussi appelé « pulse ») de 22 secondes chacun soit 220 secondes (3 minutes et 60 secondes), en laissant les échantillons au froid sur de la glace entre chaque slave de sonication. Les gels présentés correspondent à la migration des ADN à la suite de l'étape de sonication, obtenus par « reverscosslink » (dissociation des ADN et des protéines, protocole IPS2 Etape 30). Les résultats de ce test ont révélé une absence de sonication quel que soit l'échantillon considéré (Figure 29, A).

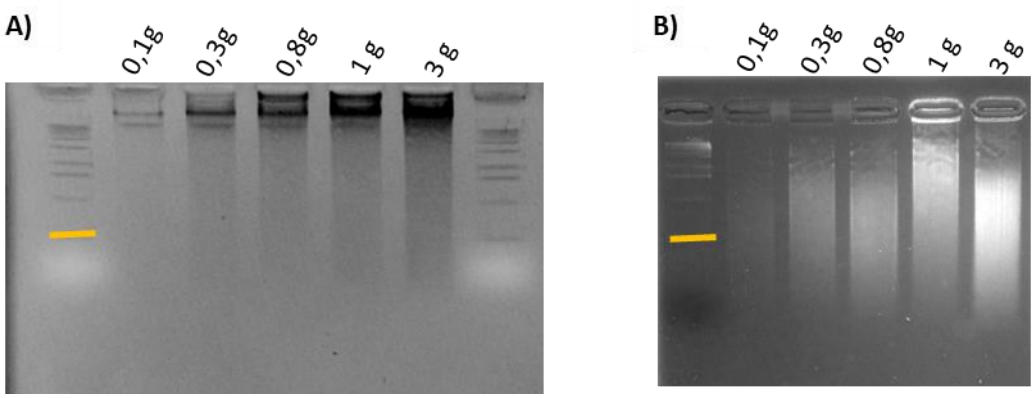


Figure 30. Test de sonication en fonction de la masse de matériel végétal récoltée (poids frais).

Les feuilles (variété Renan, pré floraison) récoltées en serres ont été directement fixées pendant 15min (formaldéhyde 1%) ; l'extraction de la chromatine a été réalisée tout de suite après la fixation ; 130µL de chromatine de chaque échantillon ont été soniqués à l'aide du covaris M220 puis reverse-crosselinked selon le protocole présenté ; 20µ de chromatine de chaque échantillon ont été déposés sur un gel d'agarose 1%.
***A)** Sonication pulsée 22 secondes de sonication, 22 secondes de pause 10 fois soit 220 secondes au total*
***B)** Sonication non pulsée 220 secondes au total*
Le trait orange symbolise 400pb d'un marqueur de taille 1kb.

En revanche, pour la sonication non pulsée, une plus importante quantité de chromatine et une meilleure sonication a été observée (B). Aujourd'hui, avec le recul, je pense que cette étape aurait mérité d'être plus étudiée car en effet, dans de nombreux protocoles, la fragmentation par sonication se fait de façon pulsée. Cependant, notons que nos temps de sonication (2 à 5 minutes) sont bien inférieurs à ceux qui avaient été utilisés à l'IPS2 pour la première expérience (10 minutes total).

Etant donné le grand nombre de tissus hétérogènes à tester pour l'expérience, nous avons ensuite tenté de déterminer les couples de temps de fixation et de sonication les plus adéquats en procédant aux tests sur deux tissus contrastés : feuilles et grains. Le grain étant un tissu très complexe, il était nécessaire d'expérimenter les différentes étapes du protocole en les comparant à un tissu plus simple pour mieux comprendre les différentes réponses au protocole de ces deux tissus très hétérogènes entre eux.

IV.2.1 **Résultats des tests fixation/fragmentation pour des feuilles matures**

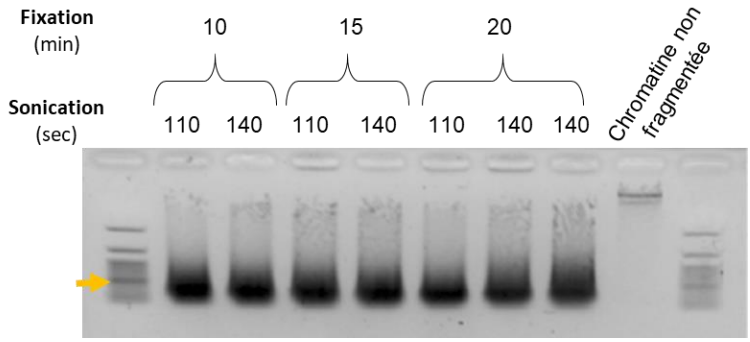


Figure 31. Gel d'agarose de migration des ADN de feuille obtenus pour différents de fixation et sonication.

Pourcentage d'agarose 1% ; échelle 100pb, Flèche orange = 500pb ; 17µL d'ADN extrait post-sonication déposé.

Sur la figure ci-contre (Figure 30), je présente un exemple de tests réalisés sur feuilles fraîches matures et vertes, prélevées au stade pré-floraison sur des plantes de la variété Renan. Nous avons testé trois temps de fixation : 10, 15 et 20 minutes et dans un premier temps pour deux temps de fragmentation : 110, 140. D'autres tests de sonication plus longue (220 et 260) sont présentés plus loin. Malgré la faible migration des ADN sur le gel présenté dans la figure 3, nous avons constaté que pour les feuilles, la majeure partie des acides nucléiques observés sont en deçà des 500pb. La fixation de 10 minutes semblait suffisante (moins de trainées observées sur les côtés du puits) comparé aux temps aux temps 15 et 20 minutes. Les temps de sonication ne semblent pas avoir d'incidence sur la taille des fragments d'ADN.

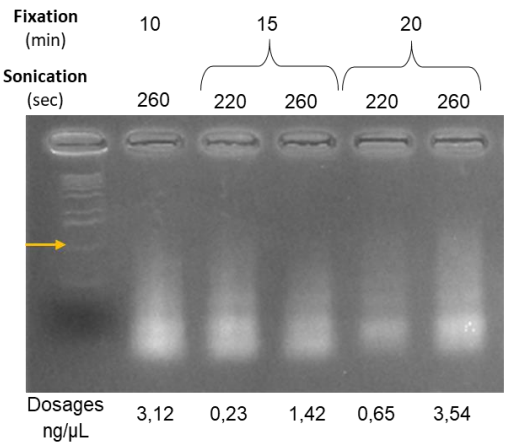


Figure 32. Gel d'Agarose de migration des ADN des feuilles obtenus après différents essais de fixation et sonication du protocole de ChIP.

Pourcentage d'agarose 1% ; échelle 1kb, flèche orange = 1000pb ; 25µL d'ADN extrait post-sonication déposé

Nous avons également testé d'autres temps de sonication. Sur le gel d'agarose de migration ci-contre (Figure 31), les smears d'ADN (profil de migration diffus des fragments d'ADN sur un gel d'agarose d'électrophorèse) des différentes conditions indiquent cette fois une faible influence des temps de fixation et de sonication sur la fragmentation des ADN. Nous nous sommes toute de même interrogés quant au degré de dégradation de la chromatine pour ces échantillons. Ici, ce sont simplement les hachures visibles dans le smear ressemble à ce que Katie Baker (ayant réalisé sa thèse sur l'optimisation du protocole chez l'Orge) avait obtenu et qualifié de chromatine dégradée dans un de ces gels (Figure 3.1 du chapitre 3 de son manuscrit de thèse).

IV.2.2 **Résultats des tests fixation/fragmentation sur des feuilles jeunes (plantules 10 jours après germination)**

Pour comparer avec les résultats ChIP-seq obtenus par l'IPS2 sur le tissu feuille de plantules de 10 jours utilisant le même protocole, nous avons également réalisé des essais de temps de fragmentation sur le

même tissu. Nous avons ainsi récolté des feuilles fraîches de plantules au stade 3 feuilles (10 jours après germination) et appliqué un temps de fixation au formaldéhyde de 15 minutes.

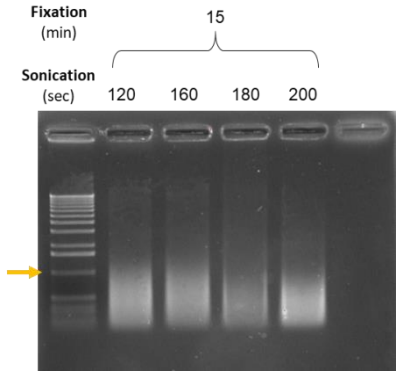


Figure 33. Gel d'agarose de migration des fragments d'ADN issus de la fixation (15minutes) de feuilles fraîches de plantules (cultivar Chinese spring) et de la sonication de la chromatine (220 secondes).

Pourcentage d'agarose 1% ; échelle 100pb, flèche orange = 500pb ; 25µL d'ADN extrait post-sonication déposé.

Le gel de la Figure 32 montre les smears d'ADN obtenus pour 120, 160, 180 et 200 secondes de fragmentation au covaris M220 (25µL déposés). La qualité des smears obtenus, quel que soit le temps de sonication indique que le protocole, en tout cas pour les premières étapes, est particulièrement adapté à ce tissu jeune, tendre, avec beaucoup de divisions cellulaires en cours.

En plus des profils de smear analysés sur gel, nous avons également beaucoup utilisé l'outil Fragment Analyser qui permet une migration de fragments d'ADN sur micro-capillaires et obtenir une courbe de distribution des tailles de fragments d'ADN présents dans un échantillon. Nous avons utilisé cet appreil pour tenter d'identifier le meilleur temps de fragmentation permettant d'obtenir une majorité de fragments d'ADN autour de 150pb, taille des fragments d'ADN enroulés autour d'un nucléosome. Pour chaque temps testé sur chacun des échantillons, nous avons utilisé le kit DNF-477 Hight Sensitivity Small Fragment Analysis Kit qui permet la migration des fragments compris entre 1pb et 1500pb.

L'annexe 3 présente les différents profils obtenus pour les ADN des tissus feuille mature utilisés précédemment. Nous pouvons constater des différences de fragmentation non décelables sur gel avec une fragmentation moins complète pour les temps 110 et 140 secondes et un pic autour de 170pb pour les sonications de 220 et 260 secondes. De la même façon, en annexe 4 sont présentés les profils de Fragment Analyser pour feuilles avec les différentes combinaisons de temps de fixation (10, 15 et 20 minutes) et de fragmentation (110 et 140 secondes). L'ensemble de ces données nous avait conduit à sélectionner, dans un premier temps le couple fixation 15 minutes/sonication 220 secondes pour les tissus de types feuille.

IV.2.3 **Résultats des tests fixation/fragmentation sur les grains (stade 500°J après anthèse)**

Pour les essais sur les grains, le problème majeur résidait dans l'accumulation de molécules d'amidon restant à la fin de l'extraction de la chromatine et donnant un culot important contenant potentiellement des contaminations de la chromatine extraite pouvant compromettre l'immunoprécipitation. Ne sachant pas comment extraire la chromatine en présence d'amidon et souhaitant identifier les potentielles pertes de chromatine dans le maillage d'amidon, nous avons dans un premier temps réalisé la sonication soit sur le culot contenant de l'amidon soit sur le surnageant, tous deux prélevés indépendamment à la fin de l'étape d'extraction de la chromatine. On constate une meilleure fragmentation pour un temps de fixation plus

court (15 au lieu de 20 minutes, Figure 33). De même, le temps de sonication semble avoir plus d'importance puisqu'on observe un smear plus bas pour le temps 140 secondes.

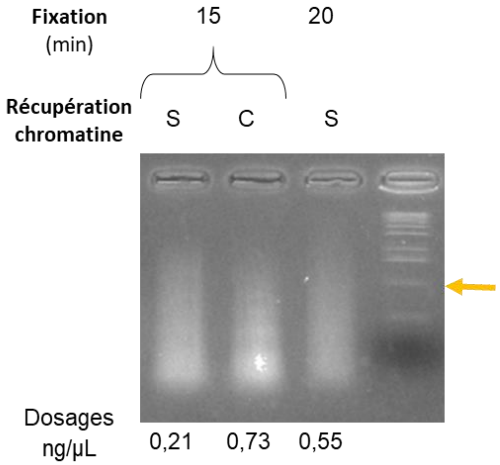


Figure 34. Gel d'Agarose de migration des ADN des grains obtenus après une sonication de 220 secondes

Pourcentage d'agarose 1% ; échelle 1kb, flèche orange = 1000pb ; 25µL d'ADN extrait post-sonication déposé. S = surnageant ; C = culot

Les profils de fragment analyser correspondant à ce gel sont présentés en annexe 5. On constate une fragmentation très partielle de la chromatine. Ainsi, les temps de 110 et 140 secondes sont potentiellement insuffisants pour ce tissu. Le profil pour une fragmentation à 220 secondes n'est pas meilleur avec plusieurs populations de fragments de différentes tailles observés. Néanmoins, il a été difficile d'interpréter ces résultats, sachant que nous avons utilisé le grain entier, constitué de plusieurs assises cellulaires et présentant cette masse d'amidon retrouvée à toutes les étapes. Ainsi, les profils de fragmentation de la chromatine pour cet organe (qui n'est pas un tissu mais un ensemble de tissus) présentent à la fois l'ensemble des fragments des différentes populations cellulaires plus ou moins actives en termes de transcription, et donc potentiellement plus de chromatine ouverte et abondante (plus facile à extraire).A la suite de ces premiers résultats, nous avons décidé de tester un protocole nous permettant de n'extraire que les noyaux des grains afin de s'affranchir des contaminants tels que l'amidon. De plus, nous pouvons constater qu'il y a de l'ADN à la fois dans le culot d'amidon et dans le surnageant pour les grains. Les dosages indiqués en bas du gel indiquent même une plus forte quantité d'ADN dans le culot contenant l'amidon. Ceci indique une perte potentille de chromatine dans le culot lors des centrifugations de l'étape d'extraction.

IV.3 Test d'un protocole d'extraction des noyaux sur grains

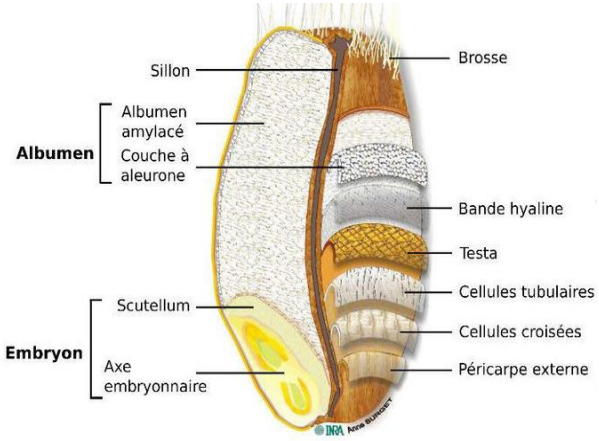


Figure 35. Schéma d'un grain de blé tendre et des tissus le composant.

Comme nous venons de le voir, les stades de développement du grain (50°J, 350°J et 500°J) sont des organes particulièrement complexes à étudier. La Figure 34 ci-contre présente l'anatomie d'un grain de blé à maturité. Ainsi, ce ne sont pas moins de 9 types de tissus ou assises cellulaires qui sont présentes dans cet organe. De plus, l'albumen amylacé représente 80 à 85% du poids total d'un grain. Les stades 350°J et 500°J

correspondent à la phase de remplissage qui s'étale de 250°J à 600°J. Parvenir à éliminer l'amidon contenu dans l'albumen au cours de l'extraction de la chromatine permettrait donc, sans dissection, de récupérer la chromatine des 8 autres couches cellulaires et notamment celle de l'embryon. C'est pour cela que nous nous sommes rapprochés d'une équipe de l'unité qui avait développé un protocole permettant d'extraire uniquement les noyaux à partir des grains. Le protocole utilisé a été publié (Bancel *et al*. 2015) et est présenté en annexe 6. Ce protocole propose une méthodologie pour extraire les noyaux en les séparant des autres contaminants ou organistes en utilisant des gradients de densité de Percoll et sucrose et des temps et puissance de centrifugation adaptés pour sélectionner uniquement les organites d'intérêt (Figure 35). Les différentes étapes théoriques d'isolement des noyaux sont les suivantes :

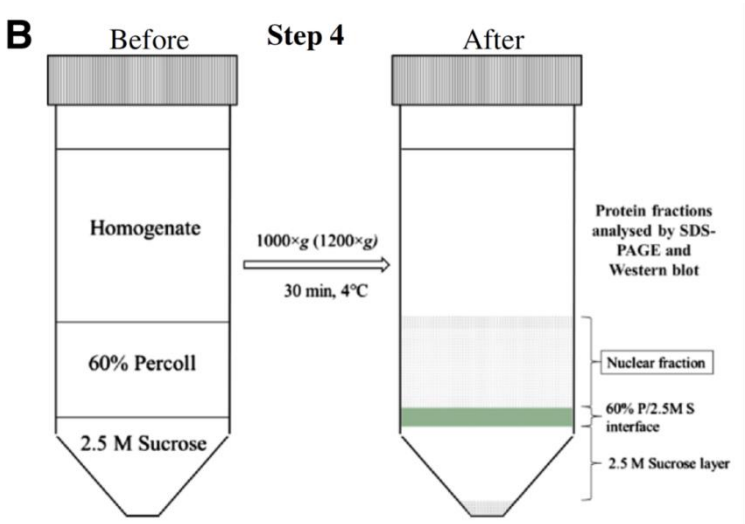


Figure 36. Schéma théorique de l'isolation de noyaux en utilisant des gradients de densité Percoll/sucrose.

- ❑ Rupture des parois et membranes cellulaires
- ❑ Filtration
- ❑ Centrifugation
- ❑ Solubilisation des membranes
- ❑ Elimination des contaminants
- ❑ Isolement des noyaux par gradient de densité et centrifugation

Après centrifugation, les noyaux sont récupérés à l'interface entre le coussin de sucrose et les protéines à l'aide d'une pipette pasteur plastique. Nous avons suivi le protocole jusqu'à la dernière centrifugation à 3500g et suspendus les noyaux dans du Nuclei Lysis Buffer (protocole IPS2).

En parallèle du test de ce protocole, nous avons également testé différents diamètres de filtration (tamis métalliques ou papier Miracloth) lors de la première phase d'extraction de la chromatine, après broyage et homogénéisation dans le tampon d'extraction 1 (protocole IPS2, étape 13). Nous avons réalisé ces tests à la fois sur des grains 500°J congelés et sur des feuilles fraîches (pré-floraison, culture sous serre, variété Renan). Nous avons repris les paramètres testés auparavant ; fixation 15 minutes au formaldéhyde 1% et sonication 220 secondes non pulsé. Le tableau 4 ci-dessous résume les tests effectués dont les résultats sont présentés sur le gel et le tableau en dessous (Figure 36). Le gel correspond à la migration des ADN à la suite de l'étape de sonication, obtenus en les décrochant des protéines par « reverscosslink » selon le protocole IPS2 (Etape 30).

Tableau 4. Plan expérimental d'optimisation du protocole de ChIP-seq pour les grains.

Protocole IPS2		Mix protocole IPS2/Bancell	Protocole Bancell
Filtration 25µm Tamis métallique	Filtration 100µm Papier Miracloth	Filtration 100µm Papier Miracloth	Filtration 25µm Papier Miracloth
Suite du protocole normale		Gradient de Percoll à la place de l'extraction buffer 3	Protocole bancel

	Filtration	Ech.	Quantité de chromatine ng/µL (Qubit)
Protocole IPS2	100µm	G_1	4.36
	25µm	G_2	6.00
Protocole Mix	100µm	G_3	0.17
Protocole Bancel	25µm	G_4	0.66
Protocole IPS2	100µm	F_1	3.36
	25µm	F_2	14.63
Protocole Mix	100µm	F_3	<0,5
Protocole Bancel	25µm	F_4	<0,5

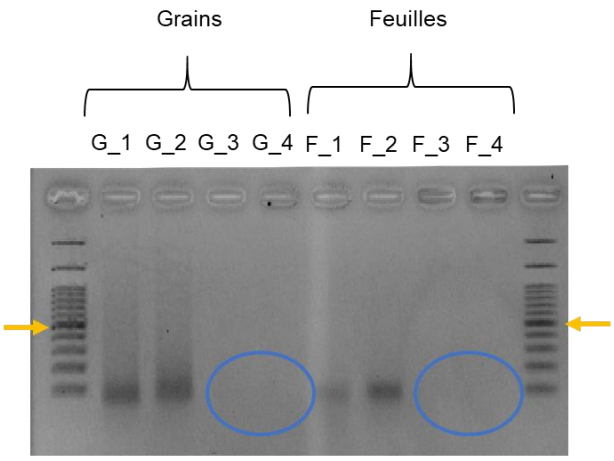


Figure 37. Résultats de dosage d'ADN et de migration sur gel d'agarose pour les essais de diamètre de filtration et protocole d'isolement des noyaux.

Pourcentage d'agarose 1% ; échelle 100pb, flèche orange = 500pb ; 25µL d'ADN extrait post-sonication déposé.

Grâce à ce test, nous avons pu mettre en évidence plusieurs choses :

- Pour le protocole IPS2, la filtration plus fine à 25µm permettait d'obtenir plus d'ADN à l'issue du « reverse crosslink », en particulier pour le tissu feuille où 4 fois plus d'ADN à été dosé (1,4 fois plus pour les grains), encadrés verts sur le tableau.
- Le protocole Bancel ou mix IPS2/Bancel ne permet pas d'obtenir suffisamment d'ADN. En particulier, ce protocole adapté pour les grains n'a pas permis de récolter de l'ADN pour les feuilles, cercles bleus sur le gel.
- Les profils de sonication constituent un amas d'ADN sur le front de migration aux alentours de 100pb et non un smear suggérant que très peu de chromatine et qui plus est dégradée a été récoltée à l'issue du protocole.

Pour résumer, ce test de protocole pour travailler de façon plus optimale sur les grains ne nous a pas paru très efficace après ce simple test et aurait mérité d'être optimisé davantage pour l'utiliser en amont du protocole de ChIP de l'IPS2. En revanche, nous avons pu constater que l'aspect filtration des débris tissulaires et cellulaires issus du broyage lors de l'étape d'extraction semblait être une étape importante pour mieux purifier la chromatine et obtenir plus d'ADN en fin de protocole.

IV.4 Comparaison des profils de sonication entre les différents tissus de la cinétique de développement

Après ces premières étapes de développement méthodologique, nous avons souhaité avoir une idée de l'ampleur de l'hétérogénéité de réponse des différents tissus à la sonication. Pour cela, nous avons arbitrairement appliqué les paramètres identifiés comme optimaux pour les feuilles fraîches matures de de la variété Renan et les jeunes feuilles fraîches de la variété Chinese Spring (stade 3 Feuilles) : fixation 15minutes et sonication 220 secondes non pulsée. Nous avons utilisé des échantillons des tissus de la cinétique congelés et surnuméraires pour effectuer le test. Nous avons appliqué le protocole de l'IPS2 en utilisant une filtration à 25µm pour la première étape d'extraction.

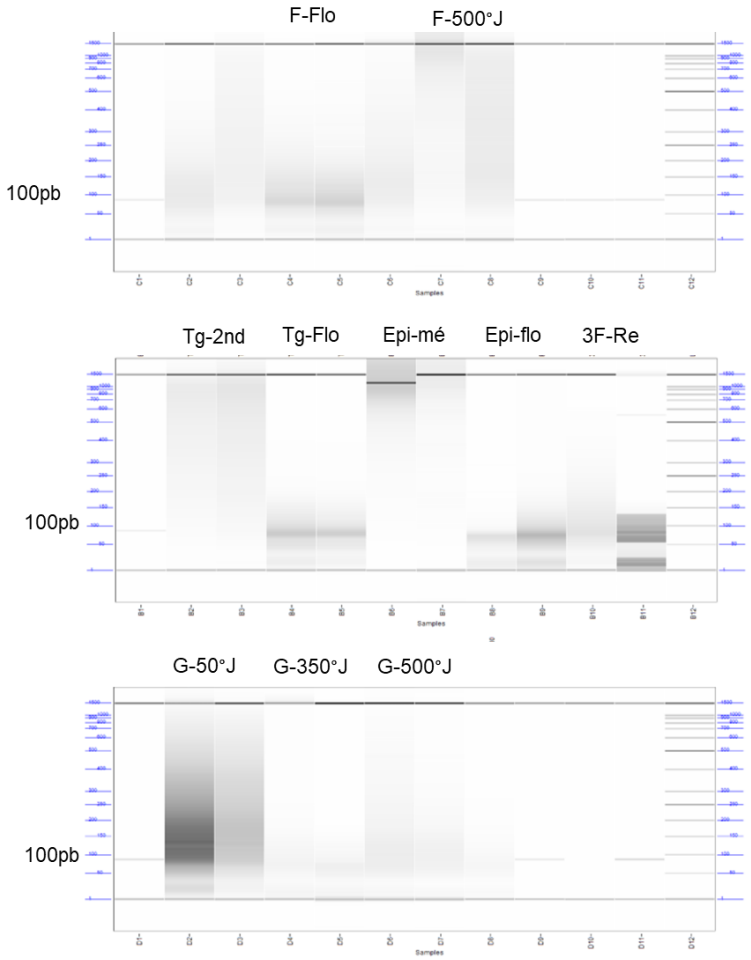


Figure 38. Profil de migration d'ADN de différents tissus de la cinétique du développement du blé tendre, fixation 15 minutes, sonication 220 secondes.

Après fixation, sonication et reverse crosslink sur deux réplicas biologiques, nous avons procédé à une migration sur Fragrment Analyser. Les gels issus de la migration sont présentés figure 37. Ces résultats indiquent deux choses :

- Une hétérogénéité de réponse entre les différents tissus avec certains trop fragmentés : Feuilles, Tige, Epis, stade floraison ; F-Flo, Tg-Flo, Epi-Flo et d'autres pas assez : Tiges au stade 2 nœuds, Epis au stade méiose, Feuilles stade 500°J après anthèse
- Des problèmes de répétabilité avec pour certains tissus une hétérogénéité de fragmentation ou de quantité d'ADN obtenue entre réplicas biologiques : Feuilles au stade 500°J, feuilles stade 3 feuilles Renan, Grains stade 50°J après anthèse, Feuilles stade 3 feuilles Chinese Srping

En termes de concentration d'ADN obtenues à l'issus de cette étape de vérification, des disparités sont aussi observables entre tissus et entre réplicas biologiques (tableau ci-contre).

Echantillons	concentration "Vérification sonication" ng/µL
Levée_1	19.8
Levée_2	37.8
3FeuillesCS_1	6.62
3FeuillesCS_2	5.6
3Feuilles Renan_1	15.1
3Feuilles Renan_2	10.8
Tige 2 nœuds_1	13.9
Tige 2 nœuds_2	47.3
Tige flo_1	9.01
Tige flo_2	10.1
Epi meiose_1	0.85
Epi meiose_2	3.2
Epi Floraison_1	21.1
Epi Floraison_2	17.8
Grains 50°J_1	4.52
Grains 50°J_2	2.09
Grains 350°J_1	3.82
Grains 350°J_2	2.7
Grains 500°J_1	4.83
Grains 500°J_2	4.55

L'ensemble de ces exemples de tests réalisés pour l'optimisation du protocole témoigne de la difficulté 1) d'internalisation et d'appropriation du protocole et 2) du développement de la technique pour étudier plusieurs tissus hétérogènes

Je présenterai dans la dernière section les résultats d'un séquençage pilote effectué après avoir réalisé le

protocole de ChIP-seq en entier sur deux tissus : feuilles Chinese spring Stade 3Feuilles et grains 500°J.

IV.5 Etude pilote : séquençage de fragments d'ADN issus d'un ChiP-seq H3K27me3

Afin d'aller jusqu'au bout du protocole, nous avions décidé de réaliser l'entièreté de l'expérience de ChIP-seq sur les deux tissus les contrastés de la cinétique de développement :

- Feuilles du stade 3Feuilles pour la variété Chinese Spring, tissus congelés
- Grains du stade 500°J après anthèse, tissus congelés.

IV.5.1 **Préparation du matériel**

Nous sommes partis de 3g de matériel de départ (poids frais congelé), en mélangeant plusieurs plantules et grains du même stade. Nous avons réalisé 15 minutes de fixation sous vide avec une solution de fixation à 1% de formaldéhyde, une fragmentation non pulsée de 220 secondes réalisée avec le Covaris M220 et une filtration à 25µm lors de l'étape d'extraction de chromatine. Nous avons utilisé l'Anticorps anti-H3K27me3 (rabbitpolyclonal, ref. 07-449, Merck Millipore, recommandé par David Latrasse de l'IPS2). 495µL de chromatine, soit 3,2µg pour les feuilles et 2,3 µg pour les grains ont été immunoprécipités avec 2,4µg d'Anticoprs. Ci-dessous sont présentés le gel de vérification de la fragmentation, les profils fragment analyser, les profils de fragments d'ADN de la banque de séquençage obtenus par la société de séquençage, les concentrations d'ADN avant et après immunoprécipitation.

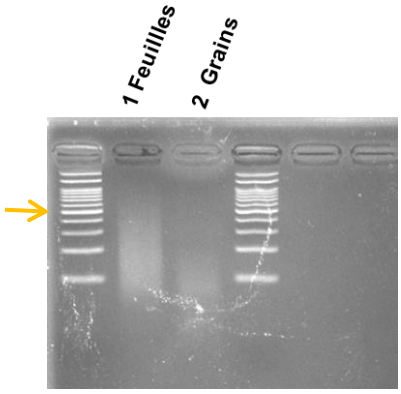


Figure 40. Vérification de la sonication de la chromatine (feuilles stades 3 feuilles et Grains stade 500°J).

Pourcentage d'agarose 1% ; échelle 1kb, flèche orange = 1000pb ; 20µL de chromatine déposée ; sonication Covaris 220 secondes non pulsé.
***1** : Feuilles stades 3 feuilles 6,62 ng/µl;*
***2** : Grains 500°J après anthèse, 4,83 ng/µl (dosage ADN Qbit).*

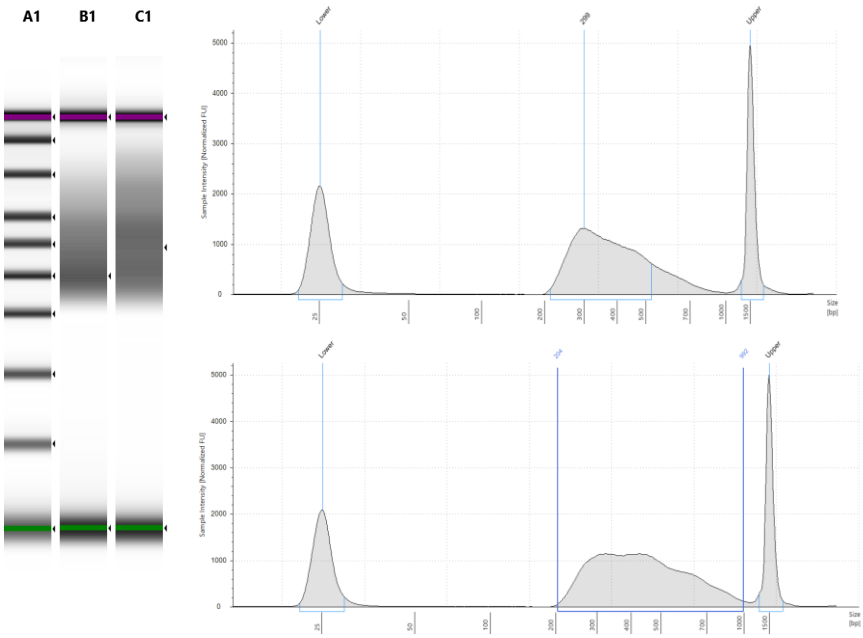


Figure 39. Profils de migration sur microcapillaires des banques de fragments d'ADN issus de ChIP-seq sur feuilles et grains chez le blé tendre.

Le gel de vérification de la sonication (Figure 39) indique une fragmentation plus importante pour l'échantillon feuille que pour celui des grains. De plus, dans de nombreux protocoles, nous avons pu lire qu'une fragmentation comprise entre 150 et 500pb était utilisée pour ensuite procéder à

l'immunoprécipitation. Cependant, pour la préparation de la banque et l'interprétation des résultats, cela doit être pris en compte car l'efficacité de PCR est plus importante pour les petits fragments (Debode *et al*. 2017). On constate sur le résultat de migration de l'entreprise de séquençage (Figure 38) une distribution plus étalée pour la banque de fragments d'ADN des grains qui comprend plus de grands fragments comparés à celle des feuilles. Ce résultat doit être pris en compte pour adapter le traitement biofinformatique des données de séquençage. En effet, des populations hétérogènes en termes de tailles de fragments d'ADN et la présence de grands fragments diminuent la résolution de la détection des sites génomiques de présence de la protéine étudiée.

IV.5.2 **Résultats**

Après immunoprécipitation, nous avons obtenu 1,4 et 0,4 ng/µl d'ADN pour les feuilles et les grains respectivement. Le rendement de l'immunoprécipitation est trois fois supérieur pour les feuilles que pour les grains avec le même anticorps alors que le différentiel de quantité de chromatine immunoprécipité n'était que de 1,4 entre les deux tissus. Nous avons ensuite procédé à la préparation de la banque de séquençage en utilisant le kit ovation (0344NB-32 Ovation Ultralow System V2) avec un nombre de cycle PCR d'amplification des fragments fixé à 14. C'est un paramètre que nous n'avons pas calibré. Une autre erreur commise a été d'envoyer au séquençage sans normaliser les banques en termes de quantité puisque la banque pour le tissu feuilles présentait une concentration de 14,2ng/µl.

d'ADN et celle des grains 1,17 ng/µl. Enfin, et parce que l'étude pilote n'était prévue que pour deux échantillons, nous n'avons pas envoyé le contrôle de type « input » correspondant à chaque échantillon ce qui est un prérequis pour pouvoir normaliser ensuite les données.

Après réception des données de séquençage nous avons procédé au pipeline habituel de traitement des données de type ChIP-seq :

1- « Trimming » des séquences qui consiste en une étape d'élimination les queues de lectures de mauvaise qualité et éventuellement des adaptateurs de la banque de séquençage s'ils n'ont pas été préalablement été enlevés (outil et options utilisés : trim_galore : --paired –phred 33 –q 30 –length 36 –illumina)
2- Alignement des lectures sur la séquence de référence Chinese spring V1.0 en utilisant l'outil d'alignement bowtie2 (options et outils utilisés Bowtie2 --very-sensitive, Samtool view -q 11 = filtration sur la qualité d'alignement (mapping quality)

Tableau 5. Sortie de terminal bowtie2 pour l'alignement des lectures de ChIP-seq H3K27me3 pour l'échantillon feuilles, stade 3Feuilles

Nombre de lectures	% sur le total de lectures séquencées	Type d'alignement
23 732 421	-	De lectures à aligner
23 732 421	100%	De lectures appariées et alignées
1 469 660	6,19%	De lectures alignées par paires 0 fois dans le génome
3 431 438	14,46%	De lectures alignées par paires 1 seule fois dans le génome
18 831 323	79,35%	De lectures alignées par paires plusieurs fois dans le génome

En exemple, je présente le résultat de la sortie brute de l'aligneur bowtie2 (Tableau 5) pour l'échantillon feuille. Ce résultat nous indiquait seulement 14,46% de lectures alignées

exactement 1 seule fois dans le génome et presque 80% avec des alignements multiples. Cela signifiait, que la banque de séquençage présentait de nombreux dupliquas (PCR ou naturels) et donc une faible complexité (diversité des fragments d'ADN). Ce paramètre est essentiel pour déterminer si l'immunoprécipitation a fonctionné correctement.

Tableau 6. Comparatif d'études de ChIP-seq étudiant la marque H3K27me3 pour les paramètres d'alignement des lectures de séquençage sur les génomes de référence.

Article	Espèce	Nombre de reads	Aligned exactly on time	Aligned more than 1 time
Zhang *et al.* 2012	Rice (seedlings)	14 412 875	9 597 200 5 (68,8%)	13 940 854
Zhou *et al.* 2016	Rice	31 817 118	21 888 363 (68,7%)	8 845 430 (27,80)
Liu *et al.* 2015	Rice	22 248 025	15 074 247 (67,76%)	6661883 (29,94%)
Zhao *et al.* 2016	Maize	18 375 337	6 929 280 (37,7%)	-

Nous avons comparé nos résultats avec ceux obtenus dans la littérature pour d'autres espèces différentes mais avec la même marque histone H3K27me3. Ces résultats indiquent des rendements bien supérieurs pour la même métrique observée (aligned concordantly exactly 1 time, Tableau 6).

Nous avons décidé de nous affranchir des dupliquas PCR en appliquant un filtre post alignement (samtools rmdup) pour les éliminer. Nous avons ensuite analysé les profondeurs de séquençage globales à l'échelle du génome puis spécifiquement sur les gènes pour chacun des échantillons (Tableau 7).

Tableau 7. Analyse des alignements pour les lectures de séquençage des tissus feuilles et grains pour un CHIP de la marque H3K27me3.

	Nombre de lectures	Profondeur de séquençage théorique	Profondeur de séquençage moyenne du génome	Profondeur de séquençage moyenne des gènes	% de gènes >20X	Profondeur moyenne des % gènes >20X	de gènes >6.5X
stade grains 500 °J	409465379	5,45 X	3.86 X	9.09 X	9.32%	27,44 X	44.30%
stade Cs3F	374174245	5 X	3.8 X	5.15 X	0.20%	50,04 X	10.00%

Très peu de gènes présentaient des profondeurs de séquençage suffisamment élevées pour les distinguer du bruit de fond. Par exemple, seuls 9% des gènes présentaient une profondeur supérieure à 20X. De plus, le calcul de la profondeur moyenne des gènes dont la profondeur était >20X est assez faible (24X). David Latrasse avait sur le même stade de développement (feuilles au stade 3feuilles) et la même marque étudiée (H3K27me3), les résultats présentés dans le tableau 8. 30% des gènes présentaient une profondeur de séquençage de plus de 30X avec une moyenne de profondeur de séquençage de 71X.

Tableau 8. Résultats de l'alignement des données de séquençage du ChIP-seq H3K27me3 réalisé par l'IPS2.

	Profondeur de séquençage moyenne des gènes	Nombre de gènes avec x>30	Couverture moyenne des gènes >30X
Feuilles stade 3Feuilles IPS2	29X	32474 (30 %)	71,22 X

Nous étant basés sur les mêmes demandes d'effort de séquençage que cette équipe, ces résultats nous ont suggéré un échec de l'expérience. J'avais procédé à un test de l'anticorps utilisé dans cette expérience par westernblot (annexe 7, protocole annexe 8). Ce western blot avait été réalisé n'avait révélé qu'une seule bande pour l'un des extraits de chromatine mais n'était pas à la taille attendue de 17kDa correspondant à

la taille des histones, comme présenté sur le western blot de la société Diagenode commercialisant l'anticorps.

Néanmoins, la comparaison de profondeur de séquençage, qui doit généralement être réalisée entre essai et contrôle (input), n'est pas la seule variable utilisée pour détecter les loci génomiques ou les protéines d'intérêt sont présentes. En effet, la symétrie des profils d'alignement des lectures entre brins d'ADN + et - constitue également un paramètre analysé par les logiciels bioinformatiques de détection de ces sites (appelés logiciel de peakcalling, pour détection des « pics » de lectures). Nous avons donc demandé à Lorenzo Concia, expert en analyses de données ChIP-seq et Hi-C de l'IPS2 de réaliser un « peakcalling » avec le logiciel MACS2 avec leur contrôle « input » séquencé, correspondant aux feuilles du stade 3 feuilles, pour le même cultivar. La Figure 40 ci-dessous représente une fenêtre de résultats d'alignement génomique des lectures de séquençage de notre échantillon feuille, de leur essai avec la même marque histone sur le même tissu, le long de 1584 kb du chromosome 2A. Ce résultat conforte notre conclusion confirme l'échec de l'expérience.

Figure 41. Comparaison des données d'alignement de CHIP-seq H3K27me3, feuilles stades 3 feuilles avec un input réalisé sur le même stade.

V. Discussion et conclusion

L'objectif de ce volet du projet de thèse était de produire des résultats de ChIP-seq pour la marque H3K27me3 sur plusieurs tissus à différent stade de développement chez le blé tendre afin de pouvoir identifier la dynamique de cette marque au cours du développement de la plante. Comme nous venons de le voir au travers de ces différents résultats, plusieurs paramètres sont à prendre en compte pour réaliser des ChIP-seq sur plusieurs tissus dont les résultats puissent être comparables entre eux. Ainsi, en ce qui concerne la comparaison de données ChIP-Seq entre tissus, le problème majeur réside dans l'hétérogénéité des tissus, qui nécessite de trouver des réponses aux questions : Comment obtenir des résultats comparables tout en prenant en compte les spécificités de chaque tissu et en particulier les paramètres fixation/fragmentation et ratio anticorps/ chromatine ?

V.1Choix des tissus

Je souhaite dans ce premier paragraphe revenir sur la sélection des tissus à comparer. Dans le tableau ci-contre (tableau 9, je résume les spécificités des tissus alors choisis pour le projet de doctorat. On constate que le nombre d'assises cellulaires par tissu est assez variable (de 3 à 9) et peuvent ainsi présenter potentiellement un état épigénétique légèrement différent (Chen et Dent 2014, Widman *et al.* 2014, Amin *et al.* 2015). De plus, certains tissus comme les grains ou les épis présentent des assises cellulaires au degré de tendreté et donc de réponse au broyage différentes mais aussi de réponses au traitement au formaldéhyde différentes. Certains tissus comme l'endosperme des grains peuvent également présenter un phénomène d'endoréduplication, qui change la ploïdie des noyaux des cellules mais aussi leur fonctionnement.

Tableau 9. Caractéristiques physio-anatomiques des tissus sélectionnés pour la thèse.

	Caractéristiques	Obstacles pour ChIP
Levée, 3Feuilles	• Tissus **jeunes**, forte division cellulaire • 4 assises cellulaires différentes	Nécessite plus de semis
Tiges	• **Tissus plus ou moins fibreux**, vaisseaux conducteurs, cambium, moelle • 5 à 6 assises cellulaires différentes	Cellules fibreuses **difficiles à broyer** **Pénétration du formaldéhyde**?
Epis méiose	• Tissus souples et faiblement chlorophylien • 7 assises cellulaires diffférentes dont une avec **meiocytes**	Ploidie et quantité d'ADN différentes
Epi floraison	• **Tissus rigides**, chrlorophyliens • 7 assises cellulaires diffférentes dont les gamètes	Ploidie et quantité d'ADN différentes **Pénétration du formaldehyde** ? **Broyage hétérogène**
Grains	• 3 stades différents, doit 3 quantités d'**amidon** et protéines de réserves différentes • **Stades physiologyques différents** (jeunes = forte division cellulaire, vieux= metaolisme spécifique= production de glucides)	**Broyage difficile** **Problème de l'Amidon** : encombrement pour l'extraction la sonication, l'IP

En effet, Polizzi *et al.* 1998 avaient évalué les densités optiques après coloration au Feulgen (qui permet de mettre en évidence les chromosomes polythéniques : plusieurs copies de chromatides) sur des noyaux de l'endosperme à différents niveaux de ploïdie chez sur *Triticum durum* cv Crezo. Ils avaient alors démontré une plus forte condensation de la chromatine des noyaux présentant une ploïdie élevée et supposaient une réduction de l'activité de transcription dans ces noyaux. Ainsi, le profil ChIP-seq d'un grain entier constitue un véritable défi nécessitant soit dissection des différents tissus le composant soit une forte robustesse statistique par la multiplication des réplicats biologiques.

Ainsi, je pense qu'il est primordial d'avoir une réflexion préalable sur la physiologie et l'anatomie des tissus pour adapter les paramètres du protocole de ChIP-seq en fonction. Je pense également que le

nombre de réplicas biologiques nécessaires doit être adapté au degré d'hétérogénéité des tissus que l'on souhaite comparer entre eux.

V.2Reproductibilité liée au plan d'échantillonnage des tissus

L'un des points essentiels que j'ai trouvé assez peu étudié par ailleurs pour les analyses ChIP-seq chez les plantes concerne la robustesse statistique de l'expérience. En effet, beaucoup de protocoles préconisent 2 à 3 répétitions biologiques. Or, plusieurs chercheurs expérimentant le ChIP-seq énoncent les problèmes de répétabilité de l'expérience dans les forums, notamment quand il s'agit de détecter des marquages différentiels entre tissus (comparativement à la découverte de nouveaux sites de liaison de marque épigénétiques ou facteurs de transcription) (exemple : https://www.biostars.org/p/274435/). Comme l'indiquent Stark et Haldfield 2016 dans leur livre dédié au design d'expérimentations ChIP-seq, la production de données de ChIP-seq est beaucoup plus sujette à variabilité par rapport à la technique RNA-seq, du fait d'un protocole plus long, avec des étapes qui peuvent présenter chacune des biais, qui se répercutent sur le résultat final. Je rajouterai que le savoir-faire de l'expérimentateur dans sa maîtrise des différents principes de biologie moléculaire (fragmentation et utilisation de machines ou d'enzymes, extraction par gradients de densité, utilisation d'anticorps et des techniques associées telles que le western blot, extraction d'ADN, technique de qPCR, préparation de banques, sans compter les compétences bioinformatiques) peut influencer les résultats de l'expérimentation. Stark et Haldfield 2016 proposent trois niveaux de réplicas :

- **Réplicas biologiques** pour capter la variance inhérente au tissu et au stade de développement et aux assises cellulaires qui le constituent (le nombre de réplicas devraient donc être proportionnels à la complexité des tissus)
- **Réplicas techniques partie laboratoire** : refaire n fois l'immunoprécipitation avec le même anticorps pour le même tissu afin de capter la variance liée à la reconnaissance antigène/anticorps
- **Réplicas techniques partie traitement bioinformatique** liés à la partie séquençage du protocole avec là aussi n re-séquençage de la banque pour capter les biais liés à l'étape de séquençage.

Grâce à l'évaluation des résultats obtenus pour les différents types de réplicas, la robustesse statistique de l'expérience pourrait être évaluée et donner plus de crédit aux résultats. L'expérimentateur pourrait alors être plus confiant pour répondre à la question : capte-t-on vraiment l'état épigénétique relatif au stade de développement ou au tissu étudié et non relatif à des biais techniques ?

V.3Optimisation des étapes de fixation/fragmentation

Comme nous avons pu le voir lors de la tentative d'optimisation des premières phases du protocole, gérer le couple fixation/sonication est une étape cruciale car ces deux paramètres influencent grandement la réussite de l'immunoprécipitation. Il est donc important de tester ces deux paramètres simultanément en étudiant systématiquement l'état de la chromatine post extraction en la déposant sur le gel de vérification de la fragmentation. Pour plus de répétabilité, je ne partirais pas de la même quantité de matériel de départ mais veillerai plutôt à prélever la quantité de matériel frais permettant d'obtenir la même quantité de

chromatine extraite pour chaque tissu. En effet, selon la tendreté du tissu et la réponse à la fixation et à l'extraction il est plus ou moins facile d'en extraire la quantité de chromatine nécessaire à l'IP.

V.4Conclusion

Ainsi, la maîtrise préalable des différentes techniques de biologie moléculaire (extraction d'ADN, extraction et fragmentation de la chromatine, utilisation d'anticoprs, quantification de concentrations d'ADN par qPCR, réalisation de banques de séquençage) sont les prérequis nécessaires pour pouvoir penser les sources de variabilité de ces différentes techniques lorsqu'elles doivent être utilisées pour des tissus différents et hétérogènes. Ces différentes sources de variations vont avoir un impact que la qualité et la quantité de fragments d'ADN finaux à envoyer au séquençage. Ces variations peuvent alors entraîner des biais et de mauvaises interprétations lors de l'analyse computationnelle. Je terminerai ma réflexion sur les différents points d'optimisation que j'ai pu relever par un commentaire sur l'analyse bioinformatique des résultats de ChIP-seq.

Au cours de ce chapitre, j'ai souhaité présenter uniquement un ensemble de résultats permettant de discuter des difficultés inhérentes au ChIP-seq que j'ai pu rencontrer et pour lesquelles je pense qu'il est important de prêter attention. En effet, depuis son développement chez l'homme en 2008, la baisse des coûts de séquençage et l'obtention de séquences de références pour un grand nombre d'espèces, cette technique c'est largement démocratisé dans les laboratoires. Or, d'après la bibliographie (articles, forums et conseils d'experts) que j'ai pu faire au cours de cette thèse, beaucoup d'ajustements en termes de méthode de fixation des protéines sur l'ADN, de méthodologie de fragmentation de l'ADN et de taille des fragments optimale, de concentrations et de qualité des anticorps utilisés (AC testés ou non sur espèces voisines, poly/monoclonaux, choix entre sensitivité et sensibilité) sont autant de sources de variations qui rendent difficile l'appropriation du protocole et la comparaison des expériences. Dans le document « Prospectives Epigénétique » publié par le CNRS en 2019 (Cahier prospective Epigénétique, écologie et évolution), les chercheurs rassemblés autour d'une réflexion concernant l'enjeux des études épigénétiques à venir ont prononcé la nécessité d'une harmonisation des protocoles d'analyse des marques moléculaires associées aux processus épigénétiques. Je cite :

« *Bien que la communauté des chercheurs en épigénétique partage les concepts, les technologies et les questions fondamentales, elle est confrontée aux difficultés liées à la multiplicité des modèles biologiques dont les connaissances scientifiques sont souvent très sommaires. Ceci nécessite la mise au point d'approches expérimentales spécifiques et différentes les unes des autres.* ***La maîtrise et la standardisation d'un éventail de techniques expérimentales sont par conséquent un des enjeux majeurs de ce domaine.*** *Des liens entre cette communauté et celles des chercheurs des domaines de la biologie moléculaire, du développement et de la reproduction, devront également être renforcés afin de faciliter les transferts de savoir et de savoirs- faire* ».

CHAPITRE V

Discussion générale

Introduction

Les travaux de recherche en génomique proposés dans ce manuscrit de thèse ont porté sur la caractérisation des biais d'expression des gènes homéologues du génome du blé tendre et des liens structure/fonction à l'échelle des chromosomes pouvant expliquer ces biais. Ces travaux s'inscrivent dans un contexte plus large visant à comprendre ce qui est sélectionné d'un point de vue génomique chez une espèce polyploïde concernant la redondance génique induite par le phénomène de polyploïdisation. Nous avons également cherché à explorer la piste de la régulation épigénétique pour expliquer les caractéristiques d'expression des gènes homéologues, notamment par l'analyse de la distribution des marques histones le long des chromosomes.

L'équipe au sein de laquelle j'ai effectué ma thèse travaille depuis 15 ans sur la caractérisation de la structure et de l'évolution du génome du blé tendre. Après avoir produit de nombreuses ressources et analyses sur le chromosome 3B telles qu'une carte physique en 2008 (Paux *et al.* 2008), l'analyse structurale et fonctionnelle en 2014 (Choulet *et al.* 2014), la caractérisation des éléments transposables (Daron *et al.* 2015), du paysage transcriptionnel (Pingault *et al.* 2015) et l'analyse des types de duplications de gènes en 2015 (Glover *et al.* 2015), l'équipe a été leader sur le séquençage du génome complet de *Triticum aestivum* cv Chinese Spring pour obtenir une séquence de référence et produire de nouvelles analyses « whole genome » (Wicker *et al.* 2018, Balfourier *et al.* 2019).

L'ensemble des résultats obtenus à travers ce doctorat permet d'apporter une vision complémentaire à ce qui avait été identifié lors de la publication de la séquence de l'ensemble des chromosomes en 2014 (IWGSC 2014). A cette époque, les principales conclusions des analyses étaient que : (i) aucune perte massive de séquences n'avait été observée (pas de fractionnement biaisé) mais légèrement moins de séquences perdues au niveau du sous génome D, (ii) plus de séquences codantes dupliquées par rapport à d'autres espèces de céréales (sorgho, orge, maïs, millet des oiseaux), présentes au niveau des régions distales des chromosomes et probablement majoritairement issues de duplications en tandem (iii) aucune dominance d'expression de l'un des sous-génomes mais plutôt une autonomie de chacun pour l'expression de ses gènes et seulement 21% des paires A-B, A-D et B-D présentant un biais d'expression, avec des spécificités selon les tissus considérés.

La caractérisation des biais d'expression pour trois catégories de groupes de gènes homéologues contrastées (triades, dyades, tétrades), reliés aux données de structure des chromosomes et aux premières données d'épigénétique sur les marques histones permettent d'apporter une vision plus précise du fonctionnement et de l'évolution de l'espace génique du blé tendre. Certaines des hypothèses de travail ont pu être traitées et sont interprétées dans la discussion qui suit.

I. Polyploïdisation et comportement transcriptionnel des gènes homéologues chez le blé tendre : tendances observées et perspectives d'études

Les *Triticeae* sont des espèces allopolyploïdes autogames, c'est-à-dire qui se reproduisent sexuellement par l'union de gamètes issues d'une même plante. Ainsi, les génomes des lignées diploïdes sont très fortement homozygotes. La polyploïdisation entraîne donc un mélange de génomes homozygotes créant une redondance génétique partielle. Or, certains de ces gènes peuvent cependant avoir évolué dans chacune des espèces progénitrices et présenter ainsi une diversité d'innovations évolutives (présence d'un paralogue, sous-fonctionnalisation épigénétique, néo-fonctionnalisation par mutation de la séquence) au sein du génome polyploïde.

L'analyse de la séquence du génome du blé tendre a permis de révéler les proportions de groupes de gènes homéologues en fonction du nombre de copies par groupe. Si l'on s'intéresse exclusivement aux gènes dits « high confidence », 51,1% des groupes sont des gènes présents en 3 copies, ratio 1 :1 :1. Cela correspond au ratio théorique attendu correspondant au degré de ploïdie pour l'espèce. En effet, le génome du blé tendre est hexaploïde et comporte trois sous génomes AADDBB. On parle de trois copies et non de six du fait du fort taux d'homozygotie entre chromosomes homologues. Ainsi, la moitié de l'espace génique de cette espèce comprend les gènes conservés chez les trois espèces diploïdes et l'autre moitié une variation du nombre de copies, avec une absence ou bien une (ou plusieurs) copie(s) surnuméraire(s) sur l'un des sous-génomes. Ces caractéristiques sont essentielles à prendre en compte pour l'analyse des biais d'expression des gènes homéologues puisque la régulation de l'expression passée (au sein des génomes diploïdes) et son devenir (au cours de l'évolution du génome polyploïde) pour ces différentes catégories de gènes homéologues peut être drastiquement différente.

Quelles sont les caractéristiques d'expression des 50% du génome présentant un ratio du nombre de copies homéologues correspondant au degré de ploïdie ?

Le premier article s'est focalisé sur l'analyse des biais d'expression pour les gènes en trois copies (1 :1 :1, triades : 1 copie sur chaque sous-génome A:B:D) pour des gènes HC et LC. Ces copies représentent les gènes les plus conservés puisqu'ils ont été maintenus au sein des génomes tout au long des 3 millions d'années de divergence et de spéciation des espèces diploïdes progénitrices (Middelton *et al.* 2014) et également à la suite des deux évènements d'hybridation aboutissant au génome du blé tendre (Marcussen *et al.* 2014). Ces gènes présentent également des positions synténiques conservées (seules 5,7% des 16 746 triades étudiées étaient non synténiques). De plus, 70% (article 1) à 80% (article 2, gènes HC) des triades ne présentent aucun biais d'expression (soit 40% des gènes high confidence du génome) et 83% d'entre elles conservent leurs niveaux d'expression relatif entre les 15 tissus étudiés (article 1). Cela signifie qu'en plus de la conservation de la séquence (identité de séquence de 95 à 99%) les niveaux et amplitudes d'expression de ces gènes sont conservés au cours de l'évolution. Aucun biais d'expression

global, spécifique à l'un des sous-génomes n'a été observé, si ce n'est une proportion légèrement plus faible de gènes « suppressed » pour le sous-génome D. Est-ce lié à des différences de niveaux d'expression entre les espèces diploïdes ou bien à la coévolution des deux sous-génomes AABB pendant environ 800 000 ans qui contraste avec l'expression du génome diploïde D ? Plusieurs éléments de réponse seront apportés à travers l'analyse des résultats obtenus dans cette discussion.

Face à ces résultats sur les triades, on peut aussi émettre l'hypothèse que ces gènes correspondraient à des gènes sensibles à l'effet dose et dont la moindre variation d'expression serait contre sélectionnée. Au sein du génome polyploïde, l'équilibre stoéchiométrique des protéines codées par les trois copies serait alors maintenu afin d'assurer la mise en place d'un phénotype adapté et dont le développement se déroulerait sans problèmes. Pour confirmer cette hypothèse, il serait intéressant d'utiliser des lignées nullisomiques / tetrasomiques ou monosmiques/trisomiques et observer si la perte de l'une des trois copies n'entraînait pas une compensation d'expression par les deux autres copies. Une étude de ce type e a été réalisée par Zhang *et al.* 2017 pour l'ensemble du génome sans distinguer les gènes en triades. Ils avaient démontré que 30% des gènes présentaient une sensibilité à l'effet dose pour les lignées aneuploïdes non-nullisomiques du blé tendre. Il serait intéressant de regarder si ces gènes sont préférentiellement des triades ou non.

Parmi ces gènes en trois copies, les analyses ont également révélé que 30% (article 1) 20% (article 2) (gènes HC) présentent des biais d'expression (triades « non-balanced »). Pour la grande majorité des groupes (14 à 25% article 1), ces biais correspondent à des niveaux d'expression relatifs plus bas pour une ou deux des copies du groupe. On pourrait alors émettre l'hypothèse d'une suppression de l'expression de certaines des copies pour retrouver une production de protéines proche de celle de l'espèce diploïde. Ces groupes de triades présentent également des biais d'expression relativement plus dynamiques au cours du développement par rapport aux triades « balanced » (62 à 73% des groupes, article 1). Ceci conforte l'hypothèse d'une potentielle sous-fonctionnalisation de certaines des copies. Par ailleurs, ces copies de gènes présentent plus fréquemment des positions relatives non-synténiques. Les biais d'expression ainsi observés pourraient alors correspondre à une régulation différentielle liée à la position génomique (environnement chromatinien différent).

Ce premier article a révélé de nouvelles tendances quant à l'expression des gènes mais n'a concerné qu'une part limitée du génome du blé tendre. Or, les triades sont en grande majorité des gènes à forts niveaux et amplitudes d'expression, associés à des protéines aux fonctions de base dans la cellule (analyses de GO) et présents dans des régions peu recombinantes du génome (régions péricentromériques), avec des vitesses évolutives en termes de mutations nucléotidiques plus lentes que les séquences présentes dans les régions distales. Ainsi, les observations et les tendances dégagées ne sont valables que pour ces gènes assez contraints évolutivement.

Quels sont les caractéristiques d'expression pour les gènes homéologues présentant des variations du nombre de copies au sein du génome polyploïde ?

A la suite du travail sur les triades, nous avons proposé d'étudier une fraction de gènes présentant un nombre de copies s'écartant du ratio théorique attendu 1:1:1 (triades). Nous avons sélectionné les gènes dits en dyades (0 :1 :1, 1 :0 :1, 1 :1 :0) représentant environ 12% des gènes « High Confidence » (HC) et les gènes présentant une copie dupliquée sur un seul des sous-génomes (2:1 :1, 1 :2:1, 1 :1 :2), environ 3% des gènes HC. Le choix de ces groupes a été fait pour prendre en compte les phénomènes de pertes de gènes et de duplication de gènes au cours de l'évolution des génomes polyploïdes. Les pertes de séquences géniques, principalement liées à une pseudigénisation non contre sélectionnée, permettent une diminution progressive de la redondance génique. La présence d'une copie dupliquée (paralogue) entraîne la diminution de la pression de sélection et la rétention de la redondance génique à travers la diversification des fonctions des différentes copies de gènes. Nous avons ainsi souhaité savoir 1) si l'absence de copies sont liées à l'évolution post-polyploïdisation ou une perte plus ancestrale et si un fractionnement biaisé est observable en s'intéressant spécifiquement aux groupes des dyades, 2) si cette perte induit une compensation d'expression par les deux autres copies, 3) dans quelles proportions les copies dupliquées précédaient l'évènement de polyploïdisation et 4) si les biais d'expression dans ces groupes de 4 gènes correspondaient à des différences liées à une évolution des deux copies paralogues chez les diploïdes ou liées à un remaniement complet des expressions à la suite de la polyploïdisation.

Les dyades

L'analyse de la conservation des variations du nombre de copies chez les espèces progénitrices nous ont permis d'estimer les variations du nombre de copies potentiellement dues aux évènements de polyploïdisation ayant conduit à l'espèce hexaploïde. Pour les dyades, les résultats ont révélé que chaque polyploïdisation a entrainé la perte d'environ 450 gènes sur chaque sous-génome qui présentent chacun environ 35 000 gènes soit une perte de 1,2%. De façon surprenante, la perte de copies sur le sous-génome D (*A. tauschii* vers *T. aestivum* : 493) équivaut à celles observées pour les sous génomes B (*T. diccocoides* vers *T. aestivum* 465) et A (498 *T. urartu* vers *T. diccoccoïdes* et 434 de *T. diccocides* vers T. aestivum). Ce résultat contraste avec ce qui a été observé à l'échelle du génome entier avec un pourcentage de gènes homéologues D retrouvés dans les groupes d'homéologie légèrement plus important pour le sous-génome D (66% vs 63% pour A et 61% pour B, IWGSC 2018). En effet, on aurait pu s'attendre à observer plus de pertes de gènes sur les sous génomes A et B puisque le premier évènement de polyploïdisation a eu lieu 800 000 avant le deuxième. Cette analyse a révélé que l'absence de séquences est majoritairement liée à la disparition d'une copie au cours de l'évolution des espèces progénitrices diploïdes. Ainsi, le phénomène de Diploïdisation Post-Polyploïdisation en est encore 1) à ces débuts ou 2) a été perturbé par la seconde hybridation avec l'espèce portant le génome D. Les chiffres énoncés quant aux variations du nombre de copies dans cette analyse sont à nuancer puisque les séquences génomiques d'une seule accession pour chaque parent et pour le polyploïde sont à l'heure actuelle disponibles. Ainsi, les polymorphismes

individuels en termes de CNV (copy number variation) ne sont pas pris en compte. Lors de notre analyse des PAV (Présence Absence Variation) entre 16 accessions de blés hexaploïdes, nous avons mis en évidence 3% de variations de présence / absence de gènes. Dans l'analyse de De Oliveira et al. 2020, sur les PAV et CNV sur les chromosomes 3B de 46 accessions, 5% des gènes du chromosome de chaque accession présentaient des PAV avec la séquence de référence Chinese Spring. Ainsi, les résultats de nos analyses sur la présence/absence des copies des dyades (mais aussi des copies dupliquées pour les tétrades) seraient à confirmer avec une analyse sur plusieurs accessions de blé hexa, tétra et diploïdes.

L'intérêt de distinguer les gènes perdus ancestralement lors de l'évolution des espèces diploïdes et ceux perdus post-polyploïdisation est d'estimer la vitesse d'évolution du phénomène de diploïdisation post-polyploïdisation ou « genome downsizing » (diminution de la taille du génome). Il serait intéressant de compléter l'analyse en estimant les proportions de pertes anté et post-polyploïdisation pour chacune des catégories de groupes d'homéologues de type 0:N :N, N :0 :N, N :N :0 (soit 489 groupes) et de confirmer ainsi les chiffres trouvés pour les dyades. Pour aller plus loin dans cette analyse, il serait également intéressant de vérifier si certains gènes homéologues considérés comme perdus ne se retrouvent pas dans les catégories pseudogènes ou fragments de gènes afin d'estimer la part de gènes en cours de pseudogénisation et ceux complètement fragmentés. Comparer les taux de pseudogénisation et fragmentation (comparaison des ratios Ka/Ks par exemple) des pseudogènes ou gènes LC en fonction de la localisation génomique des homéologues (régions distales et proximales) et des sous-génomes permettrait également d'associer ce phénomène aux caractéristiques structurales des chromosomes (taux de recombinaison notamment) et à la différence de temps d'évolution des trois sous-génomes dans le contexte polyploïde.

Ces gènes en doublet présentent des biais d'expression similaires à ceux des triades, avec une forte proportion de groupes « balanced » (64%). De plus, nous avons pu remarquer une asymétrie des biais d'expression entre les trois paires de copies AvsB, AvsD, BvsD puisque les groupes de dyades présentant des biais d'expression sont majoritairement des dyades AB. Ceci reflète ce qui a été observé pour les triades avec moins de groupes de triades présentant des gènes D « suppressed ». Cela permet de conforter deux hypothèses : 1) une expression des gènes assez similaire chez les espèces diploïdes et 2) l'hypothèse d'une potentielle initiation de l'évolution des régulations des gènes AABB après la première hybridation, au sein de l'espèce tétraploïde.

Toutefois, les limites de cette analyse sur les dyades réside dans le fait qu'il est également probable que les dyades dites « balanced » soient le reflet d'une compensation de l'expression par remaniement de la régulation de l'un des deux gènes dans le contexte polyploïde pour revenir à une dose optimale. Le principe de parcimonie voudrait cependant que ces expressions restent inchangées au cours de l'évolution des espèces diploïdes et postpolyploïdisation. A l'opposé il est également difficile de discriminer si la suppression relative de l'une des copies correspond à la nécessité d'une production de quantité de protéines précise, perturbée par la redondance génique au sein du génome polyploïde. La détermination de

la fonction de ces gènes permet d'apporter un élément de réponse. Les dyades sont très fréquemment associées à des fonctions de résistance, de stress et des fonctions de l'adressage des protéines aux protéasomes (protéines Fbox). Ces fonctions n'impliquent pas nécessairement des quantités de protéines importantes mais plutôt une expression précise et adaptée. De plus les gènes les plus réprimés (suppressed) sont les gènes homéologues des sous-génomes AB qui ont coévolué depuis 800 000 ans. Ainsi, la seconde hypothèse du retour d'une expression de ces gènes correspondant à l'état diploïde serait intéressante à explorer.

Les tétrades

La détermination de l'histoire évolutive des gènes dupliqués est cruciale pour comprendre leur évolution au sein d'un génome. Du fait d'une redondance génétique supplémentaire à celle induite par WGD, l'évolution de l'expression des gènes paralogues peut-être sensiblement affectée par l'évènement de polyploïdisation. On peut ainsi émettre l'hypothèse que l'évolution des régulations, de l'expression et du nombre de copies des gènes présentant plus de 3 copies homéologues va être plus rapide ou plus marquée post-polyploïdisation. En effet, les gènes dupliqués en tandem sont par exemple moins soumis à pression de sélection (ou sont soumis à une sélection positive) et présentent des vitesses d'évolution plus importantes que d'autres types de gènes (accumulation de mutations non synonymes) (Qiao *et al.* 2019, Guo *et al.* 2019, Panchy *et al.* 2016, Conant *et al*. 2014). Il est ainsi important d'identifier les gènes paralogues et leurs caractéristiques. Dans l'article 2, nous avons pu montrer que les gènes paralogues dans les groupes des tétrades l'étaient déjà ancestralement (chez les espèces diploïdes) pour la majorité d'entre eux puisque seulement 172 gènes dupliqués post-polyploïdisation ont été détectés au sein du génome du blé tendre. Les proportions de biais d'expression sont beaucoup plus importantes pour les tétrades avec seulement 24% des groupes présentant une expression « balanced » entre les quatre copies suggérant une potentielle distanciation fonctionnelle entre ces copies. Tout comme observé pour les triades et dyades, les biais d'expression correspondent majoritairement à la répression de l'une des copies (« suppressed ») relativement aux trois autres copies (38%). Nous avons aussi mis en évidence que les niveaux d'expression des deux copies paralogues étaient significativement plus faibles que ceux des copies des autres sous-génomes (issues de la WGD). Ainsi, les biais d'expression observés pour ces groupes s'expliquent majoritairement par les caractéristiques d'expression des deux copies ancestralement dupliquées. En effet, c'est en grande majorité (74%) une des copies des deux gènes paralogues qui est en fait réprimée (« suppressed »). Cela va plutôt à l'encontre de l'hypothèse de Freeling *et al.* 2009 stipulant que les gènes dupliqués en tandem garderaient une expression similaire (même environnement chromatinien). Or, nous avons mis en évidence que les copies paralogues correspondaient pour 64% d'entre elles à des duplications en tandem (critères de maximum 10 gènes entre deux gènes dans une distance de 1Mb de distance). Une étude plus précise des régulations de ces copies est donc à envisager pour aller plus loin dans la compréhension de leur évolution.

Les biais d'expression étaient-ils déjà présents chez les génomes diploïdes ? L'analyse des contributions d'expression à travers les 3 groupes AABD, ABBD et ABDD ont révélé que les duplications des copies A et B avaient un impact plus important sur les biais d'expression du groupe (la duplication de la copie A ayant un effet sur les biais d'expression de la copie B et vice versa). Là encore, on observe que l'histoire évolutive des gènes peut expliquer les biais d'expression au sein du génome polyploïde. Les paralogues A et B semble présenter des interrelations concernant l'expression relatives de leurs copies. Ceci pourrait aller dans le sens de l'hypothèse évoquée plus haut avec une répression de l'une des copies A ou B dans le génome tétraploïde (AABB) pour revenir à une dose de protéines stœchiométrique (2 gènes paralogues chez l'espèce diploïde ➔3 chez l'espèce tétraploïde ➔acquisition de la répression de l'une des copies). Il reste cependant à déterminer dans quelles proportions ces biais d'expression sont acquis ante ou post-polyploïdisation. Pour compléter notre analyse et répondre partiellement à cette question, nous pourrions également envisager de distinguer les 172 paires de gènes dupliqués post-polyploïdisation des paralogues dupliqués ancestralement. En effet, ces gènes-là vont présenter des spécificités d'expression plutôt liées à la polyploïdisation en particulier pour les paires AA et BB qui ont eu plus de temps pour coévoluer ensemble.

Il serait également intéressant de distinguer les gènes dupliqués en tandem des gènes dupliqués de façon dispersée dans le génome afin d'augmenter la précision de l'analyse des biais d'expression selon le mode de duplication des paralogues et de discriminer les facteurs pouvant influencer ces biais (position génomique et environnements chromatiniens différents par exemple). Cependant, les résultats dépendront des seuils (distance en Mb, nombre de gènes entre les copies) utilisés pour définir les catégories « tandem » et « dispersés » des copies paralogues.

Chez le maïs, Li *et al.* 2016 ont réalisé une analyse de réseau de coexpression des gènes en utilisant des données RNA-seq de 64 tissus distincts. Ils ont focalisé leur analyse sur les types de gènes dupliqués par WGD (whole genome doubling), dupliqués en tandem et dupliqués de façon dispersée dans le génome (inserted-duplication). Ils ont démontré que les gènes anciennement dupliqués (WGD) et ceux dupliqués sur des régions génomiques différentes (gènes dupliqués dispersés) présentent en proportion des modules de coexpression beaucoup plus divergents comparativement à des gènes dupliqués récemment et/ou en tandem. Ainsi, parvenir à décortiquer les biais d'expression liés à la polyploïdisation et ceux liés à des divergences d'expression entre espèces pro génitrices s'avère plus complexe pour les gènes présentant des paralogues qui évoluent plus rapidement. Parvenir à estimer les vitesses d'évolution des biais d'expression selon le nombre, le type et l'âge des duplications constituerait la continuité de ce travail de thèse.

Peu d'analyses bioinformatiques sont envisageables pour l'étude de l'évolution de la régulation des gènes d'un point de vue séquence nucléotidique. Les motifs d'accrochage des facteurs de transcription et leur rôle ne sont pas identifiés, notamment chez les plantes qui présentent une diversification importante de ces séquences. L'impact des TE sur la régulation des gènes n'est pas suffisamment caractérisé non plus. 25% des gènes du génome humain semblent présenter un promoteur avec une séquence dérivée d'un élément

transposable (Feschcotte 2008) et une proportion similaire a été observée pour les séquences *cis*-régulatrices au sein du génome du maïs (Zhao *et al.* 2018). Cependant, l'exacte proportion des gènes dont les séquences de régulation seraient impactées par la présence d'un TE et le rôle de ces modifications ne sont pas encore totalement compris. Chez le blé, le fait que les triades aient conservé leur niveau d'expression au cours de l'évolution des génomes diploïdes et polyploïdes malgré un renouvellement complet des TEs dans les différents génomes semble plutôt indiquer un rôle limité des TEs dans la régulation de l'expression, à cette échelle-là d'observation du moins (Wicker *et al.* 2018).

Pour expliquer ces biais d'expression, plusieurs hypothèses ont été testées : des différences d'expression déjà présentes chez les espèces pro génitrices, un différentiel de contenu en TE aux abords des gènes et un différentiel de régulation épigénétique.

Les biais d'expression entre gènes homéologues chez le blé sont-ils liés à des divergences déjà existantes chez les espèces diploïdes ?

La comparaison des biais d'expression des triades avec des données d'expression chez un génotype de blé synthétique (données RNA-seq sur 2 tissus) dans l'article 1 a révélé que 67% des triades « non-balanced » présentaient des biais d'expression différents d'avec le génotype de blé synthétique hexaploïde et 41% de ces derniers présenteraient des expressions transgressives par rapport aux espèces pro génitrices, surtout pour le sous génome D. Cette analyse suggère ainsi que certains des biais d'expression observés pour les triades seraient dus à des divergences déjà présentes entre espèces pro génitrices. Les résultats de comparaison avec le génotype synthétique sont à nuancer car les biais d'expression observés peuvent être induits par des réarrangements chromosomiques et les effets immédiats de l'hybridation. De plus, les données RNA-seq utilisées dans ces analyses n'ont pas été produites exactement dans les mêmes conditions. Il serait donc intéressant de réaliser ces comparaisons avec plusieurs individus synthétiques dont les ARN seraient extraits avec le même protocole pour s'affranchir des biais techniques.

La comparaison des données d'expression des dyades et tétrades avec les données chez des espèces tétraploïdes et diploïdes n'a pas été proposée dans l'article 2 et constitue une perspective de travail intéressante.

A ce titre, une analyse plus poussée des dynamiques évolutives de l'expression des gènes homéologues chez le blé tendre nécessiterait un système évolutif complet, du même type que celui utilisé par Edger *et al.* 2017 pour *Mimuls* et Song *et al.* 2017 pour le coton. En effet, l'absence de données de transcription des gènes homéologues obtenues simultanément sur plusieurs tissus pour un système biologique comprenant les trois espèces pro génitrices *Triticum urartu*, *Triticum speltoïdes* et *Aegilops tauschii*, l'hybride F1, le génotype polyploïde synthétique et l'espèce polyploïde actuelle *Triticum aestivum* ne permet pas, pour l'instant, d'explorer rigoureusement l'impact des hybridations et de la polyploïdisation sur l'expression des gènes chez le blé tendre.

Dans cet objectif, il serait également intéressant d'opter pour des analyses dont la normalisation des données RNA-seq permette de tenir compte des niveaux de transcriptions correspondant au degré de ploïdie des espèces considérées. En effet, l'étude de Coate and Doyle 2010 a révélé que la quantité d'ARNm total de l'espèce tétraploïde *G. dolicocarpa* était 1,4 fois supérieure à ses pro géniteurs. De la même manière, Visger *et al.* 2017 ont mis en évidence des différences de détection des régions différentiellement transcrites entre l'espèce diploïde et l'espèce autotétraploïde du genre *Tolmiea*. Ces différences étaient liées au type de normalisation des données RNA-seq faite soit 1) par transcriptome ou 2) par cellules, la première étant basée sur la concentration et ne prenant pas en compte les différences de taille de transcriptome entre espèces. Ils ont mis en évidence que l'espèce autotetraploïde tend à conserver une abondance de transcrits par unité de biomasse proche de l'espèce diploïde en augmentant la quantité de transcrits par cellule, notamment pour des loci reliés à des fonctions de photosynthèse.

De plus, les données RNA-seq sont normalisées de façon classique en divisant l'abondance des lectures de séquençage a un locus donné par le total des lectures alignées sur le génome (Transcripts per millions of reads) supposant que la quantité d'ARN extraite est égale dans chaque échantillon, ce qui est rarement le cas (Chen *et al.* 2016). Ceci équivaut à une normalisation par transciptome et donc par concentration de transcrits. Un facteur de correction peut être appliqué en pondérant les abondances de lectures ainsi normalisées par un facteur correspondant au degré de ploïdie. Par exemple, dans l'article 1 du manuscrit, les données RNA-seq issues du tétraploïde ont été pondérées par 2/3 et les données issues de l'individu synthétique par 1/3 pour pouvoir comparer les expressions des gènes avec l'espèce hexaploïde. Or, si les tailles de transcriptomes varient selon le degré de ploïdie, les gènes identifiés comme différentiellement exprimés ne sont pas nécessairement exprimés à des niveaux différents mais peuvent au contraire être maintenus à des concentrations différentes, qui dépendent de la taille du transcriptome (Visger *et al.* 2017). Il serait alors intéressant de se pencher sur la question des biais techniques liés à l'identification des gènes différentiellement exprimés entre espèces de différentes ploïdies avant d'étudier spécifiquement les questions relatives aux changements d'expression des gènes homéologues.

La structuration des chromosomes peut-elle expliquer les caractéristiques d'expression des gènes homéologues ?

La structure des chromosomes du blé est connue pour présenter une spécificité propre aux *Triticeae* avec des régions distales dynamiques en termes de délétion de séquence (notamment TE) et de taux de recombinaison et des régions proximales dépourvues d'évènements de recombinaisons et enrichies en séquences répétées (Choulet *et al.* 2014). Ces caractéristiques font que les séquences vont évoluer de façon plus rapide dans les régions distales. Cela peut expliquer en partie l'évolution différentielle des gènes homéologues selon les régions génomiques : si l'une des copies homéologues est touchée par un évènement de recombinaison endommageant la séquence, l'expression des autres copies peut potentiellement changer suite à cette délétion pour retrouver un niveau de protéine efficace. Cette hypothèse pourrait en partie expliquer ce que nous avons observé, à savoir que les biais d'expression, qui

concernent toutes les catégories de groupes d'homéologues triades, tétrades, dyades, sont observés préférentiellement pour les groupes présents dans les régions distales des chromosomes. En effet, les triades présentant un biais d'expression sont préférentiellement enrichies dans les régions distales et les tétrades et dyades sont des gènes majoritairement retrouvés dans ces mêmes régions. Enfin, dans l'article 1, les biais d'expression des triades ont été comparés entre deux accessions de blé tendre (Chinese spring et Ayrvana). Cette analyse a révélé qu'une forte proportion de groupes de triades des régions proximales présente une conservation des biais d'expression entre cultivars (85%), alors que les groupes des régions distales présentent eux plus de variations des biais d'expression. Ainsi, les gènes sont distribués de façon non aléatoire le long des chromosomes en fonction de leurs caractéristiques d'expression et de leur fonction. Cependant, nous avons mis en évidence dans l'article 2 que les triades présentent systématiquement des niveaux d'expression plus importants que les tétrades et dyades quel que soit le compartiment chromosomique. De plus, les fonctions des gènes en dyades et tétrades conservent des fonctions d'adaptation quel que soit la région génomique. Ces données témoignent des limites de l'hypothèse structurale liée à la séquence pour expliquer le fonctionnement du génome.

Pour expliquer ces spécificités d'expression des gènes homéologues conservées entre espèces mais pouvant varier pour certaines entre accessions et entre régions génomiques, l'hypothèse de l'implication de la régulation épigénétique est proposée. Comme cela a été observé chez d'autres espèces, la répartition des environnements chromatiniens et des éléments transposables le long des chromosomes sont des facteurs pouvant expliquer les caractéristiques d'expression des gènes homéologues (Makarevich *et al.* 2013, Baker *et al.* 2015, Roudier *et al.* 2011, Song et al 2017, …).

II. Qu'est-ce que l'analyse de la régulation épigénétique permet d'expliquer sur l'expression des gènes homéologues ?

Les mécanismes épigénétiques sont à l'origine de l'expression différentielle et spécifique des gènes faisant que des cellules comportant les mêmes séquences d'ADN vont présenter des expressions différentes des gènes. Ce processus permet en effet la mise en place de cellules différenciées, présentant un ensemble de protéines spécifiques, qui vont s'agencer entre elles pour former des tissus puis des organes. De même plusieurs études ont démontré que les processus épigénétiques interviennent également dans l'adaptabilité des individus aux aléas environnementaux avec une adaptation de l'expression des gènes corrélée aux contraintes du milieu (Fernandez *et al.* 2014, Baulcombe et Dean 2014, Schmid *et al.* 2018, Thiebaut *et al.* 2019). L'exemple du fonctionnement du locus FLC témoigne de la plasticité de l'expression des gènes impliqués dans la floraison en fonction des contraintes environnementales (Whittacker et Dean 2017). Les éléments ou mécanismes épigénétiques précis intervenant dans ces processus et qui seraient soumis à pression de sélection ne sont pas encore complètement connus. Cependant, l'hypothèse du rôle potentiel de ces mécanismes dans la stabilisation des génomes néo-polyploïdes et dans l'évolution de la redondance génique post-polyploïdisation est actuellement à l'étude. En effet, les environnements chromatiniens définis par les combinaisons de marques épigénétiques (méthylation de l'ADN, modifications post-

traductionnelles des histones) évoluent-ils post-polyploïdisation pour s'harmoniser ? Les environnements en TE aux abords des gènes sont-ils remaniés ? L'étude de ces phénomènes constitue un intérêt majeur pour l'étude des espèces néo-polyploïdes. En effet, les changements épigénétiques, identifiés pour être dynamiques (au cours du développement) et réactifs (aux variations environnementales), peuvent constituer les premiers changements progressifs et identifiables de l'évolution d'un génome néo-polyploïde.

Les biais d'expression entre gènes homéologues chez le blé sont-ils liés à des différences de contenu en TE aux abords des gènes ?

Concernant l'implication des éléments transposables, deux hypothèses sont à l'heure actuelle étudiées. Les variations de présence/absence ou de densité de TE au niveau des promoteurs ou dans les 1500pb en amont des gènes, déjà présentes chez les espèces diploïdes pro génitrices ou provoquées par transposition de TE par dérégulation du silencing post-polyploïdisation peuvent entraîner :

1) Des modifications de la séquence et donc des sites d'interactions des facteurs de transcription/éléments de régulation et autres protéines se liant à l'ADN et intervenant dans la régulation de l'expression des gènes
2) Des modifications de l'environnement chromatinien aux abords des gènes avec étalement de la méthylation de l'ADN des TE vers les promoteurs selon leur proximité, ce qui donne un environnement chromatinien défavorable à la transcription et conduit au silencing des gènes

Comme déjà observé chez d'autres espèces, les divergences de contenu en TE aux abords des gènes entre espèces pro génitrices sont souvent corrélées à des biais d'expression au sein du polyploïde (Song *et al.* 2017, Edger *et al*. 2017). Or, dans l'article 1, pour les triades « non-balanced », aucune différence de contenu en TE au niveau des 1500pb en amont des gènes n'a été détectée. Ainsi, ces biais d'expression observés ne semblent pas être corrélés à des différences de contenu en TE dans les régions régulatrices. Si la méthylation de l'ADN au niveau des promoteurs et séquences régulatrices en amont des gènes n'a pas été étudiée, en revanche, il a été mis en évidence des variations de densité de méthylation CG dans le corps des gènes avec un gradient de densité décroissant des gènes « Balanced », « Suppressed » et « Dominant ». La méthylation des cytosines dans le corps des gènes a été associée à une expression constitutive des gènes de ménage.

Si aucune différence n'a été trouvée concernant les TEs au niveau des séquences régulatrices pour expliquer les biais d'expression des triades, ce paramètre semble être relié à la dynamique transcriptionnelle des triades à travers les différents tissus et stades de développement. Pour les 10% de triades pour lesquelles le biais d'expression du groupe étaient très différents en fonction des stades considérés (triades « dynamiques »), ces variations sont corrélées à des divergences de séquence en termes de proportion de TE dans les régions régulatrices proches des gènes (sur 1,5kb en amont des gènes). Ainsi, selon ces résultats, la transposition de TE aux abords des gènes n'expliquerait pas les biais mais plutôt les

sous-fonctionnalisation tissus-spécifique. Ces dynamiques d'expression pourraient correspondre à 1) des sous-fonctionnalisation des gènes pour leur expression tissu spécifique chez les espèces diploïdes ce qui induit un biais chez le polyploïde ou 2) l'acquisition et la sélection d'une sous-fonctionnalisation post-polyploïdisation. Si tel est le cas (1), on devrait observer des divergences de contenu en TE entre les espèces diploïdes. Or, ces différences de proportion de TE en amont des triades dynamiques (1500pb) concernent 20% d'entre elles (soit environ 10 000 gènes) et sont de l'ordre de 8% (88% triades dynamiques vs 79% triades stables). Le fait que cette analyse soit faite sur des triades aux caractéristiques transcriptionnelles extrêmes en termes de variation de biais d'expression au cours du développement et que les différences observées ne soient que de l'ordre de 8% met en avant les limites de l'analyse. De plus, la comparaison du contenu en TE des trois sous-génomes du blé tendre a montré que ces séquences ont subi un turnover (i.e. remplacement par cycles d'amplification/perte) depuis la divergence des espèces diploïdes il y a ~3 millions d'années mais que le contenu en TE entre les trois espèces pro génitrices du blé tendre reste le même. L'hypothèse de l'implication des TEs dans les différences de régulation des gènes homéologues chez le blé n'est donc pas entièrement vérifiée et reste à explorer. Par ailleurs, un différentiel d'identité de séquence entre les triades dynamiques et les triades stables (séquences codantes et protéiques) a révélé des ratios de Ka/Ks plus élevés pour les triades dynamiques. Ainsi, les différences d'expression des gènes entre tissus et stades de développement pourraient aussi être expliquées par une évolution de leur fonction (néo-fonctionnalisation) par divergence de séquence plutôt qu'une divergence de régulation liée aux différences de contenu en TE.

Concernant les différences de biais d'expression relatifs entre les paires de gènes homéologues des triades « non-balanced » A vs B et A vs D et B vs D, l'analyse de la table S6 de l'article 1 permet de constater que les sous-génomes A et B présentent une proportion de groupes dits « suppressed » plus élevée par rapport au sous-génome D. A contrario, les proportions de groupes d'homéologues avec l'une des copies définie comme dominante sont très similaires pour les 3 sous-génomes. De plus, à travers les 15 tissus les copies D sont beaucoup moins fréquemment associées à la catégorie « suppressed » comparativement aux copies A et B. Les proportions des TE, comparées entre paires de triades A vs B, Avs D et B vs D, n'ont pas été conduites pour ces triades. Dans l'article de Wicker *et al.* 2018, des biais d'enrichissement en familles de TE aux abords des gènes (2kb en amont) entre sous-génomes comparativement au reste du génome ont été observés. Une investigation plus claire des biais de proportion de TE, et de méthylation de l'ADN de ces régions, en amont des gènes en fonction des sous-génomes (AB vs AD et vs BD) pour les groupes de gènes présentant un biais d'expression permettrait de clairement répondre à la question de l'implication de l'environnement en TE pour la sous-fonctionnalisation des gènes ante et postpolyploïdisation. En effet, certains des réarrangements dans l'environnement direct des gènes ont pu être sélectionnés car induisant une sous-fonctionnalisation avantageuse, mais peuvent être non significatifs ou indécelables si l'on n'étudie pas spécifiquement les gènes avec biais d'expression.

Afin de pouvoir compléter les analyses réalisées dans l'article 1 sur les triades, il serait intéressant de réaliser des comparaisons de divergence de proportion de TE au niveau des 1500pb en amont des gènes

homéologues également pour les catégories tétrades, dyades. Pour les tétrades, l'analyse de ce paramètre pour les gènes paralogues serait également très prometteuse. En effet, il est assez surprenant de voir que les gènes présentant préférentiellement des biais d'expression importants (dyades, tétrades) sont situés dans les régions du génome où les TEs sont plus fréquemment éliminés (Daron *et al.* 2015). D'ailleurs les distances inter géniques dans ce compartiment chromosomique sont beaucoup plus contraintes que celles au niveau des régions proximales (Wicker *et al.* 2018). Ainsi, le lien entre proportion de TE et distance entre gènes pourraient amener à considérer la régulation des gènes non plus en termes linéaire mais plutôt en 3D. Ces paramètres pourraient être compris comme pouvant influencer l'agencement des gènes dans l'espace nucléique et dans des boucles de régulation (favorables ou défavorables à la transcription) comme décrites dans Concia *et al.* 2019.

L'analyse des patrons de méthylation des cytosines entre les sous-génomes ABD du blé tendre par traitement BS-seq n'ont pas permis, à l'échelle du génome, de révéler des divergences de méthylation de l'ADN entre des trois sous-génomes (Gardiner *et al.* 2015). Cela reste cohérent avec le fait que le contenu en TE entre les trois espèces pro génitrices du blé tendre est sensiblement le même ; en effet, la méthylation de l'ADN est principalement observée sur les séquences répétées et est impliquée dans le silencing de ces dernières. En revanche, cela contraste avec ce qui a pu être observé pour d'autres allopolyploïdes plus anciens tels que le coton ou le maïs (Song *et al.* 2017, Zhao *et al.* 2017) pour lesquels les contenus en TE, notamment aux abords des gènes sont différents entre les espèces pro génitrices et corrélés à des différences de méthylation de l'ADN. Ainsi, plusieurs éléments contradictoires sur l'implication de l'environnement en TE sur l'expression des gènes restent à être confrontés pour trancher sur cette hypothèse. On peut également se demander si l'implication des différents acteurs de l'épigénétique ne dépend pas du compartiment chromosomique considéré.

La répartition des états chromatiniens peut-elle expliquer les spécificités d'expression des gènes homéologues chez le blé tendre ?

Pour aller plus loin dans l'exploration de l'hypothèse épigénétique, l'implication des marques histones pour expliquer les spécificités d'expression des gènes homéologues chez le blé tendre est une autre piste qui a été explorée au cours de ce doctorat. Plusieurs hypothèses étaient envisageables : 1) le maintien ou l'acquisition d'un environnement épigénétique favorable à la transcription permettrait de conserver l'équilibre de dose qui doit être maintenu pour les gènes sensibles à l'effet dose, 2) l'acquisition d'un marquage épigénétique de répression transitoire de la transcription permettrait la sous-fonctionnalisation de l'une des copies comme l'implication de la marque H3K27me3 dans la sous-fonctionnalisation des gènes dupliqués au sein d'un génome polyploïde comme déjà certaines étudié par Berke *et al.* 2012, Berke *et al.* 2014, Arthur *et al.* 2014, Zhu *et al.* 2017, Chica *et al.* 2017.

La production des données de ChIP-seq pour les marques H3K27me3, H3K36me3, H3K9ac et H3K4me3 sur l'un des tissus de la cinétique de développement (feuilles stade 3 feuilles) lors de la publication de la séquence génomique de référence du blé tendre (IWGSC 2018) a permis en partie d'explorer les

hypothèses précitées. Dans l'article de l'IWGSC, il avait été mis en évidence que la distribution de ces marques épigénétiques le long des chromosomes présente un partitionnement spécifique avec un enrichissement des marques associées à une transcription des gènes active (H3K36me3, H3K9ac et H3K4me3) au niveau des régions proximales des chromosomes et un enrichissement de la marque de l'hétérochromatine facultative (H3K27me3) au niveau des régions distales.

Concernant les triades, les groupes « balanced » présentent un enrichissement des marques H3K9ac, H3K36me3 et H3K4me3. Ceci a été confirmé par les résultats de Li *et al.* 2019. Par ailleurs, dans l'article 2, nous avons montré que les triades présentent en grande majorité un marquage identique entre les trois copies du groupe (64.4%) avec en grande majorité la marque H3K9ac (88,5%). Ces gènes sont donc associés à un environnement chromatinien similaire entre les trois sous génomes. Cela témoigne du rôle prépondérant de ces gènes dans les fonctions de base de la cellule puisqu'ils sont associés à un environnement chromatinien favorisant leur transcription active, constitutive et à des niveaux comparables. Ces gènes sont localisés dans les ces régions proximales, faiblement recombinantes sont donc propices à une stabilité d'expression avec peu d'évènements de recombinaison pouvant entraîner des remaniements d'environnements chromatiniens. On peut donc imaginer que ce compartiment génomique semble être favorable à l'expression des gènes sensibles à l'effet dose, nécessitant une synchronisation stable de leur expression.

En opposition à ces gènes, les triades non-synténiques et « non-balanced », les dyades et les tétrades sont des groupes d'homéologues qui présentent des expressions plus spécifiques (amplitude moyenne d'expression de trois conditions au lieu de douze pour la marque H3K9ac) et des niveaux plus faibles et plus fréquemment associés à la marque H3K27me3. Les densités de cette marque sur les gènes sont également plus importantes pour les gènes « non-balanced » (groupes d'homéologues avec un biais d'expression) des tétrades et des dyades. Les gènes en tétrades et dyades présentent également une expression plus dynamique au cours du développement. En effet, la proportion de groupes de dyades et tétrades présentant une diversité d'expression des copies dans les différents organes est plus importante que pour les triades (55, 79 et 35% respectivement). Les gènes des dyades et tétrades présentent également moins de consistance dans leur marquage au sein des groupes, par rapport aux triades, avec plus de groupes présentant l'un des homéologues marqué différemment comparativement aux autres (pour les deux marques H3K9ac et H3K27me3 étudiées) Ces résultats confirment le rôle de cette marque dans la spécificité de la répression de l'expression des gènes et une expression spécifique (tissus/stades de développement).

Les ratios de proportion de gènes marqués H3K27me3 et H3K9ac sont pratiquement inversés pour les dyades et les tétrades lorsqu'on les calcule dans les régions distales et régions proximales. Cependant, nous avons également observé que, quelle que soit la région chromosomique, les triades sont plus fréquemment associées à la marque H3K9ac. Il y a donc bien une relation entre marquage épigénétique et partitionnement des gènes selon leurs fonctions et caractéristiques d'expression.

Cependant il est encore difficile de savoir si la sélection naturelle a favorisé l'agencement des environnements chromatiniens sur de larges régions chromosomiques, permettant un partitionnement fonctionnel chromosomique, ou bien selon les caractéristiques fonctionnelles de chaque gène à une échelle plus locale avec un environnement épigénétique qui lui est propre.

En regard des résultats de l'article 2, on peut émettre l'hypothèse que la sélection naturelle a pu favoriser la localisation de gènes en fonction de comment (dans quels tissus, à quels moments, dans quelle quantité) ils doivent être exprimés conférant aux chromosomes un partitionnement particulier. Pour apporter des éléments supplémentaires dans cette caractérisation épigénétique des gènes homéologues, il serait intéressant dans un premier temps de comparer les différences de cohérence de marquage des gènes homéologues des différents groupes dyades, tétrades, triades selon les régions chromosomiques.

Cette question peut être également analysée à travers l'étude de la conformation tridimensionnelle de la chromatine. Chez le blé, une étude de ce type a été récemment réalisée sur des cellules méristématiques de plantules de 14 jours a été réalisée et a permis d'identifier des patrons de distribution des marques épigénétiques d'un point de vue tridimensionnel (Concia *et al*. 2020). En cherchant les interactions locales entre gènes, les auteurs ont défini un modèle théorique avec des régions condensées (appelées ICONs Intergenic CONdense Spacer) définissant des boucles aux extrémités desquelles des gènes activement transcrits se retrouvent en contact. Ils ont corrélé ces domaines avec des densités des différentes marques épigénétiques (Figure 41).

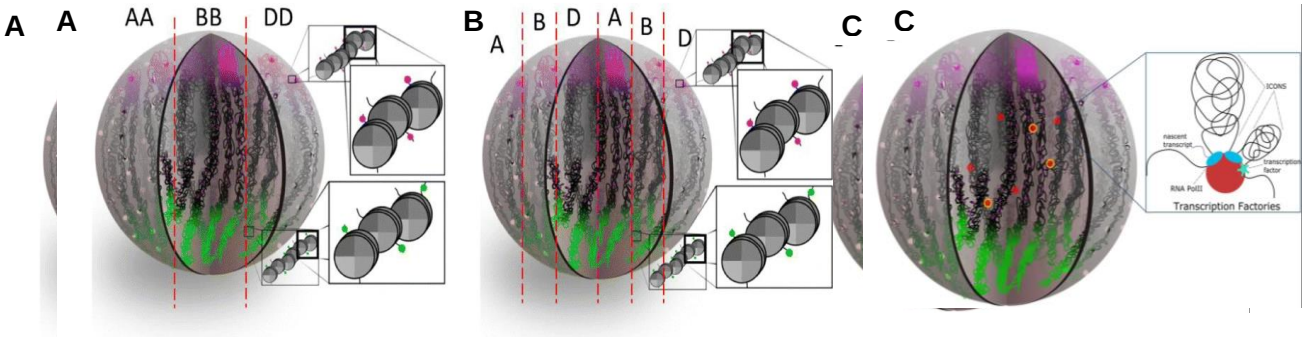


Figure 42. Conformation de la chromatine dans le noyau de cellules de blé tendre.

***A)** et **B)** Schéma théorique de la répartition potentielle des chromosomes homéologues au sein du noyau et distribution des marques épigénétiques de l'euchromatine ou hétérochromatine facultative (vert H3K27me3) et de l'hétérochromatine (rose, H3K27me1). **C)** Schéma théorique des boucles d'interactions entres gènes définissant des domaines d'interaction et de coexpression des gènes. Concia et al. 2020.*

Les marques histones balisant les zones activement transcrites se situent aux frontières des ICONS, là où les gènes sont accessibles à la machinerie de la transcription alors que les régions repliées des ICONS sont plutôt associées à la marque H3K27me3. Ces résultats très innovants ouvrent la voie à l'analyse du fonctionnement et de l'évolution des génomes d'un point de vue de la dynamique de la conformation (forme) de la chromatine et notamment comment et à quel point les épis génomes d'un génome polyploïde s'intègrent et s'harmonisent.

Les états chromatiniens peuvent-ils évoluer et participer à l'évolution des génomes polyploïdes ?

Dans un génome polyploïde, la concordance des états chromatiniens entre gènes et régions génomiques est cruciale pour assurer la synchronisation de la production de protéines, notamment pour des gènes codant des protéines dont l'équilibre stœchiométrique est essentiel pour former des complexes moléculaires (gènes sensibles à l'effet dose). Or, nous avons vu que l'une des hypothèses pouvant expliquer la perte de séquences pour retourner vers un état diploïde réside dans l'acquisition de marquages épigénétiques défavorables à l'expression entraînant une plus faible pression de sélection et la pseudogénisation de séquences (Bottani *et al.* 2018). De même, la sous-fonctionnalisation des gènes peut également se faire *via* des remaniements épigénétiques de type marque histones (Wang *et al.* 2014, Wang *et al.* 2016, Wang *et al.* 2017, Xu *et al.* 2018). On peut donc s'attendre à voir une évolution du marquage épigénétique au sein d'un génome polyploïde. Dans l'article 2, nous avons remarqué que la proportion de gènes marqués H3K9ac suit un accroissement des tétrades AABD vers les tétrades ABDD et l'inverse a été observé pour la marque H3K27me3. Les gènes A et B des tétrades AABD et ABBD sont des gènes qui ont davantage co-évolué que les tétrades ABDD ; on peut donc émettre l'hypothèse que cette coévolution découle ou a provoqué une évolution du marquage épigénétique post-polyploïdisation. En particulier, nous avons remarqué que les gènes paralogues du groupe des tétrades étaient significativement moins associés à la marque H3K9ac. De plus, nous avons également mis en évidence dans les deux articles que, quelle que soit la catégorie de groupe d'homéoCNV considérée, les gènes du sous-génome D sont mois réprimés relativement aux autres copies et qu'ils sont moins associés au marquage H3K27me3. Ceci pourrait témoigner de l'intégration récente du sous génome D dans le contexte allo polyploïde, qui n'aurait pas encore entrainé beaucoup de changements épigénétiques tels que l'acquisition de ma marque H3K27me3.

Même si ces résultats ne sont valables que pour un seul tissu, on pourrait se demander si l'acquisition progressive de la marque H3K27me3 ne serait pas impliquée dans la sous-fonctionnalisation de l'une des copies des paralogues. Pour étudier cette question, il peut être intéressant de comparer la conservation du marquage épigénétique entre espèces apparentées qui représentent une estimation de l'évolution de leur marquage épigénétique.

Nous avons réalisé cela avec deux espèces de céréales pour lesquelles nous avons pu trouver des données de ChIP-seq pouvant être comparées avec celles du blé en termes de marques et de tissus/stade de développement : le riz et le maïs. Nous avons observé que s'il existe un différentiel de conservation de séquence entre les gènes conservés en triades et les gènes évoluant plus rapidement (dyades et tétrades) entre les trois espèces, le marquage épigénétique semblait lui na pas diverger drastiquement. En effet, les gènes des différents groupes présentaient les mêmes proportions de conservation de marquage entre les trois espèces. En revanche, nous avons observé une plus grande conservation de gènes marqués H3K9ac que les gènes marqués H3K27me3. Ceci pourrait conforter l'hypothèse de l'implication de cette marque dans l'évolution de la régulation des séquences et de l'expression des gènes homéologues. Ces résultats restent cependant à nuancer car ils ne concernent qu'un seul tissu de développement et n'ont pas été produits dans les mêmes conditions expérimentales. Il est donc primordial d'obtenir des données de ChIP-seq sur un nombre de tissus plus important afin de discriminer les processus développementaux des

processus évolutifs, bien que ces derniers soient intimement liés selon le courant de pensée de l'Evo-Devo.

L'ensemble des résultats sur le marquage épigénétique des gènes homéologues chez le blé tendre présentés dans ces deux articles semble conforter l'hypothèse d'une implication des états chromatiniens dans les caractéristiques et l'évolution de l'expression de gènes homéologues. En particulier, la marque H3K27me3 qui confère une expression spécifique des gènes semble toucher les gènes à évolution rapide et les groupes d'homéologues ayant co-évolué plus longtemps entre eux (AABB). L'étude proposée par Berke *et al.*2012 sur 3169 paires de gènes paralogues marqués H3K27me3 chez *Arabidopsis thaliana* avait démontré que deux paralogues marqués H3K27me3 présentaient peu de divergence d'expression et de divergence de séquence au niveau des régions régulatrices mais des taux les plus élevés de divergence de séquences codantes (corps des gènes) en comparaison avec les paralogues non marqués H3K27me3 ou lorsque seulement un des deux paralogues était marqué. En revanche, ces paires de paralogues ayant seulement un des deux gènes marqués H3K27me3 présentaient des divergences d'expression et de séquences régulatrices très marquées. Ainsi, la présence de la marque contraindrait l'évolution des gènes paralogues. Le lien entre épigénétique et évolution de la redondance génique semble complexe. Existe-il un lien entre régulation épigénétique et potentiel d'évolution des séquences ? Cette étude montre également qu'il est crucial d'étudier en parallèle les divergences de séquence et les divergences de marquage épigénétique pour ne pas conclure à l'implication unique des variations épigénétiques quant à l'évolution des variations d'expression des gènes dupliqués.

Ces considérations montrent qu'une analyse précise des biais de marquage épigénétique entre gènes homéologues en les séparant selon des catégories évolutives (gènes dupliqués paralogues ou gènes ayant peu d'ortho logues), de localisations chromosomiques (par régions ou territoires au niveau des chromosomes) ou fonctionnelles (selon les biais et caractéristiques d'expression) permettrait de mieux identifier leurs spécificités épigénétiques et de mieux les caractériser.

Enfin, les analyses des deux articles ont été réalisées en utilisant des données de ChIP-seq ayant été produites sur un seul tissu et avec un seul réplica biologique. Même si les résultats concernant la détection des pics de marques semblent être confortés par les données proposées par Li *et al*. 2019, il est difficile de démêler les relations épigénétiques et expression des gènes sans avoir des données sur différents tissus. En effet, la marque H3K27me3 est associée au destin cellulaire et à la différentiation des cellules par répression progressive et transitoire de la transcription des gènes au cours du développement (Hennig et Derkacheva 2009). Ainsi, comment distinguer les patrons de marquage épigénétique associés à l'expression des gènes au cours du développement de ceux associés à l'évolution de la régulation de gènes homéologues en n'utilisant qu'un seul tissu ?

C'est bien la nécessité d'obtenir des données sur plusieurs tissus qui avaient été mise en avant lors de la conception de ce projet de doctorat et il avait été proposé de produire des données de ChIP-seq sur plusieurs tissus provenant d'une cinétique de développement du blé tendre.

III. Produire des résultats de ChIP-seq, perspectives d'optimisations.

Une analyse comparative de marques épigénétiques inter-tissus nécessite de connaître les éléments pouvant impacter la reproductibilité et la robustesse de l'expérience afin de pouvoir calibrer le protocole sur chacun des tissus tout en ayant la possibilité de comparer les résultats entre eux. Parmi ces paramètres figurent les propriétés de la chromatine en elle-même, les propriétés des tissus étudiés ainsi que le type de marque épigénétiques étudiées notamment en termes d'abondance dans le génome (Zhao *et al.* 2020, Li *et al.* 2014, Landt *et al.* 2012). La fixation des tissus, la fragmentation de la chromatine, l'immunoprécipitation et le choix de la profondeur de séquençage représentent les différentes étapes pour lesquelles il convient de minimiser l'hétérogénéité entre réplicas biologiques et entre échantillons.

Reproductibilité du ChIP-seq : préparation des tissus

Le projet de développement du ChIP-seq du projet doctoral a révélé l'hétérogénéité de réponse entre tissus pour les premières étapes notamment pour la fixation et l'extraction de la chromatine. Une grande variabilité en termes de concentrations d'ADN obtenues en fin de protocole et des profils de fragmentation très différents ont en effet été observés (chapitre IV).

La première étape pouvant induire un biais lors de l'analyse de plusieurs tissus simultanément est la fixation des interactions protéines/ADN au formaldéhyde (cross-linking). Selon les tissus végétaux que l'on souhaite étudier, la pénétration du formaldéhyde peut sensiblement varier. Pour pallier ce problème l'une des solutions à envisager consisterait à réduire en poudre les tissus au froid (glace + azote) et de fixer ensuite les tissus au formaldéhyde à température ambiante comme suggéré dans plusieurs études (Salvic *et al.* 2013, Vimont *et al.* 2019). Cette technique permettrait d'augmenter l'homogénéité de l'action du formaldéhyde et d'éliminer les variations de fixations liées aux propriétés imperméables de certains tissus (feuilles ou tiges avec cuticules cireuses par exemple). Cependant, l'hétérogénéité de réponse au broyage entre tissus doit être maîtrisée. Chez les souris, certains auteurs avaient mis en place un protocole de hachage des tissus pour favoriser la pénétration du formaldéhyde (Schmidt *et al.* 2009). Cependant, cette technique présente l'inconvénient d'introduire une autre forme d'hétérogénéité : la taille des morceaux hachés, qui peut différer selon les échantillons. Des tests sur l'efficacité et l'homogénéité de fixation (cross-linking) sur des tissus hétérogènes en utilisant la méthode de fixation sur les tissus réduits en poudre constitue donc une approche intéressante dans le cadre de la réalisation d'un atlas de ChIP-seq de tissus hétérogènes. Le calibrage du temps et des conditions de fixation des interactions protéines/ADN au formaldéhyde constitue une étape cruciale pour la reproductibilité de l'expérience dans le but d'obtenir des quantités de chromatine similaires et représentant des interactions ADN/protéines spécifiques du tissu.

Il serait intéressant de vérifier l'homogénéité de la fixation en utilisant un contrôle permettant d'estimer le taux de fixation entre échantillons. Pour cela, Baranello *et al.* 2016 ont par exemple utilisé des cellules humaines transformées avec soit 1) le système hétérologue GFP (ne se liant pas à l'ADN) fusionné à un peptide d'adressage au noyau soit 2) soit un peptide d'adressage fusionné à la protéine Top1 se liant aux

promoteurs de gènes transcrits. Ils ont évalué la quantité d'ADN obtenue après immunoprécipitation par qPCR et évalué le degré de fixation de ces protéines en fonction de différents temps d'incubation au formaldéhyde. Les tests qPCR après immunoprécipitation avec des anticorps anti-GFP et anti-Top1 sur des séquences connues ont révélé que des fixations prolongées de formaldéhyde (60 minutes vs 4 ou 10 minutes) entraînent une fixation non spécifique de protéines à l'ADN. En effet, des fragments d'ADN ont été retrouvés pour les cellules présentant le système hétérologue GFP. Dans leur étude, ils sont aussi mis en évidence qu'un différentiel entre 25 et 37°C de température de tampon de fixation entraîne également des différences de rendement en ADN après IP. Un tel système, utilisé pour comparer le différentiel de fixation selon les échantillons permettrait d'estimer le delta de reproductibilité de l'étape de fixation en fonction de la réponse des différents tissus au broyage.

Reproductibilité du ChIP-seq : fragmentation de la chromatine

Lors de la phase d'optimisation du protocole de ChIP-seq, nous avons pu remarquer une forte hétérogénéité dans la réponse des différents tissus à l'étape de fragmentation. Une fixation au formaldéhyde des différents tissus inappropriée, une extraction de la chromatine plus compliquée pour certains tissus aux parois rigides par exemple, peuvent avoir un impact sur cette étape. Ainsi, les temps de fragmentation sont à optimiser selon les deux paramètres précédents. Le succès de l'immunoprécipitation, la complexité de la banque de séquençage et donc la résolution de la détection des pics d'enrichissement de lectures en fonction du ratio signal/bruit de fond dépendront particulièrement de l'étape de fragmentation.

La chromatine est une structure dynamique et hétérogène : les zones fermées, ou hétérochromatine, présentent des marques épigénétiques (types marques histones) relativement stables et une structure très dense. En revanche, la chromatine dite active ou ouverte, l'euchromatine, est beaucoup plus dynamique (transcription, remodelage) avec des facteurs de transcriptions, des ARN polymérases ainsi que nombre important de protéines se liant à l'ADN viennent interagir avec cette dernière. Elle présente également des combinaisons de marques épigénétiques complexes. Les études de la conformation 3D de la chromatine permettent de définir des topographies de cette entité au sein du noyau (appelés compartiments chromatiniens) et montrent les différents niveaux d'hétérogénéité de sa compaction et de son organisation. Cette hétérogénéité est importante à prendre en compte puisque les zones d'hétérochromatine constitutive ou d'hétérochromatine facultative sont moins facilement fragmentées comparativement aux zones d'euchromatine, ce qui va conditionner l'accessibilité des épitopes à l'anticorps et la purification des fragments d'ADN en fin d'expérience. Ce biais entre régions d'hétérochromatine et d'euchromatine peut induire une surreprésentation de cette dernière dans les fragments séquencés et la détection de faux positifs. A l'inverse, pour l'hétérochromatine qui est moins bien fragmentée, les enrichissements faibles pourront être éliminés si le bruit de fond est trop important (Chen *et al.* 2012). D'un point de vue théorique, la fragmentation optimale de la chromatine devrait correspondre à des fragments de la taille d'un nucléosome qui permet l'empaquetage de 147pb d'ADN. Ainsi, pour l'analyse des marques histones,

si la majorité des fragments de chromatine correspond à la taille d'un nucléosome, l'alignement des lectures obtenues à l'issue du ChIP-seq permettra un positionnement plus précis des pics d'enrichissement de lectures de séquençage correspondant à la présence de la marque histone d'intérêt, lors de l'étape de peak-calling. Or, la résolution des expériences actuelles de ChIP-seq ne permettent pas encore d'obtenir une détection de marquage d'une résolution équivalant au nucléosome (Zhang *et al.* 2019). En effet, de longs fragments peuvent être associés à plusieurs nucléosomes qui ne portent pas forcément les mêmes modifications épigénétiques.

Il existe deux façons de fragmenter la chromatine : l'utilisation d'une endo et exonucléase, la MNAse et l'exposition aux ultrasons (sonication). Le tableau résume les avantages et inconvénients des deux techniques. Selon certains auteurs, la fragmentation par une enzyme de type MNAse serait contrindiquée lorsque les échantillons sont fixés au formaldéhyde car cela réduit l'accessibilité de l'enzyme aux sites de restriction et donc son efficacité (Haring *et al.* 2007, Scientific support : www.abcam.com/technical, Das *et al.* 2004). La comparaison de différentes méthodes de fragmentation de la chromatine de Skene et Henikoff (2015) pour l'étude du positionnement de la PolII sur des cellules S2 de Drosophiles (Schneider cells) a démontré que la précision de la détermination de la position génomique de la polymérase dépendait de la taille des fragments obtenus lors de l'étape de fragmentation, qui elle-même dépendait de la densité de nucléosomes au niveau des promoteurs plus ou moins « actifs ». En conclusion de leur étude, ils proposent une optimisation de l'étape de fragmentation de la chromatine : une première étape de fragmentation par ultrasons suivie d'une digestion enzymatique. Lors du contrôle de la qualité de la chromatine sur un gel d'agarose, une quantité trop importante de très courts fragments d'ADN doit indiquer une fixation trop faible (Ricardi *et al.*2010) ou une fragmentation trop importante pouvant abîmer les épitopes reconnus par les anticorps.

Tableau 10. Avantages et inconvénients de deux moyens de fragmentation de la chromatine.

	Avantages	Inconvénients
MNAse	Préservation des épitopes	Biais dans l'action de l'enzyme selon les séquences ADN
	Pas cher	L'accessibilité de certains sites de restriction peut varier selon les types cellulaires et les organismes – Contre indiqué avec crosslinking.
	Digestion jusqu'aux extrémités des nucléosomes	Selon les lots, les activités enzymatiques peuvent varier et induire des biais entre échantillons
	Permet de conserver les interactions lorsque la chromatine n'est pas fixée	Importance de pondérer les quantités d'enzymes selon la quantité de chromatine à digérer
		Facteurs de non-reproductibilité : température, présence de sels, concentrations
Sonication	Pas de biais liés à la séquence	Biais lié à la conformation : hétérochromatine moins bien fragmentée
	Plus de reproductibilité entre échantillons	La présence de SDS et de mousse peut altérer la sonication
	Permet de lyser aussi des cellules non encore lysées et de fragmenter plus d'ADN	Possibilité de reverse crosslink ou de changement de conformation des protéines si la sonication est trop forte en présence de SDS

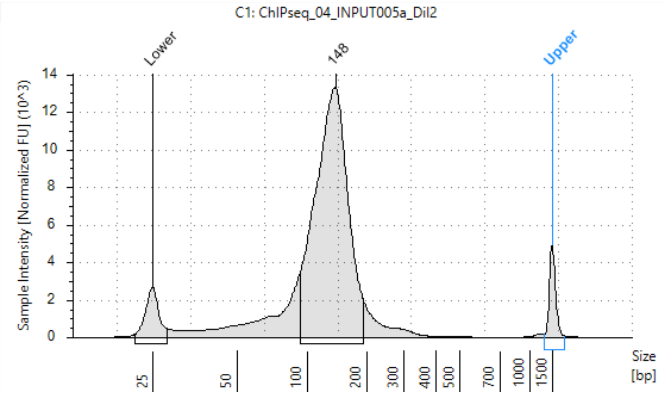


Figure 43. Distribution des fragments d'ADN fragmentés à la MNAse.
Tissus feuilles stade 3 feuilles ; fixation 25 min (formaldéhyde 1%) ; traitement à la MNAse (200U) pendant 12min.

Face aux difficultés rencontrées au laboratoire par rapport à l'utilisation de la fragmentation, des tests sont en cours au laboratoire quant à l'utilisation de la MNAse. La Figure 43 montre que, pour le contrôle input, la distribution des fragments obtenus par traitement de la chromatine à la MNAse sur feuilles de plantules donne des résultats prometteurs.

Reproductibilité du ChIP-seq : immunoprécipitation

Dans le projet de thèse, deux études pilotes avec immunoprécipitation et séquençage de fragments d'ADN issus des IP ont été réalisées sur le stade trois feuilles (deux fois) et sur le stade grain 500°J. Les résultats ont révélé une absence d'enrichissement par rapport à un contrôle « input ». Un test de l'anticorps en western blot n'avait pas été positif mais les conditions de western blot étaient-elles suffisamment optimisées ?

Le choix de l'anticorps et le calibrage de l'immunoprécipitation ciblant la protéine ou la modification épigénétique d'intérêt est le second point le plus important pour la réussite de l'expérience de ChIP-seq. Ainsi, l'optimisation des tests des anticorps en western blot doit également être pensée dans un projet de ChIP-seq afin de valider au préalable les anticorps sur l'espèce étudiée. Egalement, des tests qPCR sur des régions connues du génome peuvent être réalisés pour vérifier l'enrichissement spécifique lié à l'immunoprécipitation. Or, il est nécessaire de connaître au préalable des gènes spécifiquement marqués par la marque épigénétique d'intérêt. Ceci est difficile à évaluer pour la marque H3K27me3 qui est une marque dynamique au cours du développement. De plus, les rendements en ADN sont si faibles à l'issue du protocole que les tests qPCR ne correspondent qu'à une très faible partie du génome (Baker 2015).

Le succès de l'immunoprécipitation repose à la fois sur 1) le degré de similarité tridimensionnelle de l'épitope d'intérêt avec celui à partir duquel l'anticorps a été synthétisé 2) l'accessibilité de l'épitope pour l'anticorps et 3) le ratio de concentration adéquat entre chromatine et anticorps.

La spécificité et la sensibilité des anticorps ne sont pas maîtrisées par l'expérimentateur mais sont testées lors du processus de fabrication par les entreprises. La sensibilité correspond à la capacité de l'anticorps à s'associer avec plusieurs sites portant le même épitope au sein du génome. La spécificité correspond à la fiabilité de l'anticorps pour son association avec l'épitope d'intérêt et non un épitope proche dans le génome (exemple des marques histones H3K27me2 et H3K27me3). Pour obtenir des résultats de ChIP robustes et de qualité, il serait préférable de sélectionner des anticorps qui ont été testés en western blot

sur des protéines en conditions natives et non-dénaturantes. Comme mentionné dans l'article définissant les consignes à suivre pour réaliser des ChIP-seq dans la recherche épigénétique chez l'humain (projet ENCODE, Landt *et al.* 2012), il est recommandé de tester les anticorps en western blot avant de les utiliser pour vérifier les tests opérés par les entreprises de fabrication. En effet, les anticorps sont testés sur des espèces modèles et sur des tissus spécifiques (cf protocoles ChIP-seq Abcam, Diagenode, Merx Millipore) qui ne correspondent pas forcément à l'espèce et au tissu d'intérêt.

Le type de production, anticorps poly ou monoclonaux peut également avoir une incidence sur la qualité de l'immunoprécipitation. En effet, des anticorps polyclonaux possèdent une sensibilité plus élevée car ils reconnaissent potentiellement plus d'épitope dans l'extrait de chromatine. Cependant, leur spécificité peut être problématique notamment pour des marques histones dont seules un groupement méthyle diffère entre deux épitopes (Ex H3K27me3 ou H3K27me2). *A contrario,* les anticorps monoclonaux sont plus spécifiques mais souvent disponibles en de plus faibles quantités dans le commerce. Ainsi, choisir des anticorps monoclonaux assure une plus grande spécificité mais une moins bonne sensibilité (moins de signal) et vice versa pour les anticorps polyclonaux. Or, lors d'une expérience de ChIP-seq sur plusieurs tissus il est important de penser à l'homogénéité des lots lorsqu'un même anticorps doit être utilisé sur plusieurs échantillons, afin d'éviter des biais liés aux spécificités des lots.

En ce qui concerne la quantité d'anticorps à utiliser pour réaliser l'immunoprécipitation, ce paramètre n'est pas fixé universellement et nécessite également d'être testé dans les conditions d'expérimentation du laboratoire. En effet, les concentrations conseillées par les fabricants ont été testées en western blot et ne correspondent donc pas aux conditions de ChIP de l'expérience que l'on réalise. Pourtant, de la concentration optimale en anticorps par rapport à la concentration en chromatine va dépendre le *ratio* signal/bruit de fond qui est critique pour la résolution de l'expérience. Des indications d'ordres de grandeur sont conseillées telles que 1 à 10µg d'AC pour 25µg de chromatine https://docs.abcam.com/pdf/chromatin/A-beginners-guide-to-ChIP.pdf https://www.abcam.com/protocols/cross-linking-chromatin-immunoprecipitation-x-chip-protocol).

Cependant, la façon la plus optimale de définir la quantité d'anticorps à utiliser serait bien sûr de tester soi-même ce paramètre dans des conditions les plus proches de celles de l'expérience finale. En effet, tester ce paramètre sur un échantillon ayant une quantité de chromatine bien inférieure à celle de l'expérience finale peut donner une indication biaisée puisque la visibilité des épitopes peut être différente selon la concentration de la chromatine. L'idéal serait donc de trouver la concentration pour laquelle la quantité d'anticorps arrive à saturation dans la reconnaissance de ses cibles sans la dépasser, ceci pouvant conduire à une aspécificité réduisant le *ratio* signal/bruit de fond. Un test qPCR sur les sites de fixation de la protéine étudiée (ou de l'histone avec une modification spécifique), connus en termes de séquence pour pouvoir créer des amorces, pourrait alors être réalisé pour plusieurs ratios de concentrations [anticorps] / [chromatine] sur chacun des tissus.

Etat de l'art des analyses ChIP-seq sur plusieurs tissus chez les triticeae

Lors de son doctorat, Katie Baker (Baker 2015) a travaillé sur l'optimisation de protocoles de ChIP-seq pour obtenir des données sur l'orge pour différentes marques histones (H3K4me3, H3K9me1, H3K36me3, H3K9me3). Cette espèce est une espèce diploïde appartenant à la famille des Triticeae et donc proche du blé. Elle a testé trois protocoles de ChIP-seq publiés et tenté d'optimiser les différentes étapes de fixation, fragmentation par ultrasons, immunoprécipitation pour son projet. Après avoir modifié et optimisé un nombre important de paramètres, elle n'a obtenu que très rarement un enrichissement spécifique pour les différentes marques histones qu'elle avait sélectionné. En effet, les profils d'enrichissement en contrôle qPCR entre les différentes marques testées étaient très similaires. Après ces tests sur des protocoles de laboratoire, elle a utilisé un kit commercial grâce auquel elle a pu obtenir des résultats corrects. Cette thèse met en avant le fait que l'optimisation du ChIP-seq est également très complexe pour différentes marques histones mais sur un seul tissu, Le challenge d'un projet inverse (tel que celui prévu initialement pour mon doctorat) : plusieurs tissus pour une seule marque est tout aussi important voire plus difficile car les kits commerciaux n'ont, pour la plupart, été développés que pour un ou deux tissus spécifiques.

Très peu d'études ChIP-seq sont publiées pour plusieurs tissus chez les plantes. L'analyse des données utilisées pour la création de la Plant Chromatin DataBase par Liu *et al.* 2018 illustre bien le fait que pour l'instant, les principaux tissus utilisés pour réaliser des études de ChIP-seq chez les plantes sont des tissus jeunes de type plantules. On peut citer l'étude pionnière de Makarevich *et al.* 2013 chez le maïs qui avait étudié H3K27me3 sur 4 tissus différents : l'embryon, les fleurs mâles et femelles et l'endosperme du grain en réalisant trois réplicas biologiques. Bien que cette analyse ait été réalisée par un ChIP suivi d'une hybridation sur puce à ADN (ChIP on ChIP), peu de détails sont donnés quant à l'optimisation du protocole concernant les différents tissus (plantule, panicule, fleur mâle, endosperme, embryon) dans la section matériel et méthode de l'article.

Pour l'instant, ce sont des comparaisons de tissus deux à deux qui sont réalisées, comme cette étude très récente sur une comparaison de données de ChIP-seq H3K27me3 entre les inflorescences et les feuilles chez *Brassica rapa* réalisée par Paya-Milans en décembre 2019 en utilisant deux réplicas biologiques pour les feuilles et un pour l'inflorescence. Dans leur article, ils ont comparé les résultats de deux logiciels d'alignements de lectures de séquençage et deux logiciels de peak-calling pour valider la reproductibilité de l'analyse bioinformatique. Ils ont également réalisé des analyses de corrélation entre réplicas biologiques et entre contrôles (Inputs) et réplicas biologiques. Pour les deux réplicas feuilles, la corrélation entre les pics obtenus pour les deux échantillons est de 94%. Les coefficients de corrélation entre inputs et réplicas biologiques qui traduisent la reproductibilité de l'expérience est de de 81% pour le couple Input_feuille_1/input_Infloresence et de 72% pour le couple Input_feuille_2/input_Infloresence. Ainsi, un effort est proposé pour la reproductibilité du traitement bio-informatique du ChIP-seq, qui elle aussi dépend de l'optimisation de certains paramètres (profondeur de séquençage notamment Landt *et al.* 2012,

Chen *et al.* 2012), mais la reproductibilité de l'expérience en elle-même n'est pas réellement testée ou validée.

Chez le blé, deux jeux de données sont disponibles à l'heure actuelle pour des expériences de ChIP-seq sur les marques histones : IWGSC 2018 et Li *et al.* 2019, réalisées sur le même tissu et stade de développement. Les auteurs de la dernière étude ont comparé leurs résultats à ceux publiés précédemment en calculant des coefficients de corrélation concernant la densité de lectures aux pics d'enrichissement communs détectés dans les deux études. Les coefficients de corrélation sont en moyenne de 80%, le plus faible étant trouvé pour H3K27me3 (71%) (Figure 43 page suivante). Cette étude, qui a été réalisée sur le même tissu confirme la reproductibilité de l'expérience pour ce tissu, même si le nombre de pics retrouvés en commun entre les deux études n'est pas mentionné.

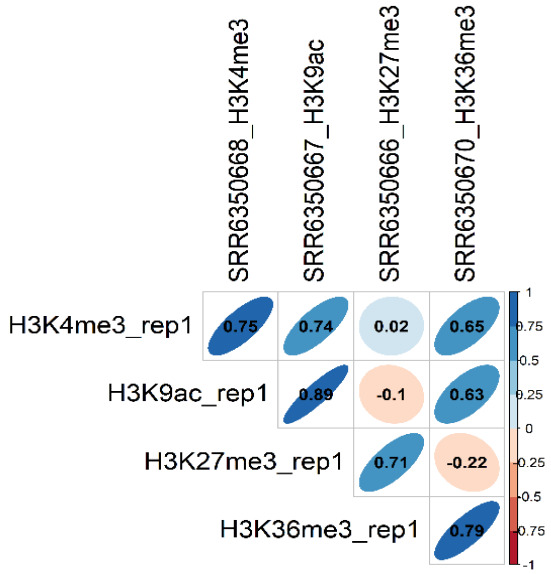


Figure 44. Diagramme des corrélations de Pearson des densités de lectures entre deux expériences de ChIP-seq H3K27me3 chez le blé tendre

Les densités de lecture ont été calculées sur des fenêtres de 500pb.et les coefficients de corrélation sur les fenêtres se chevauchant entre les deux études IWGSC 2018 et Li et al.. 2018.

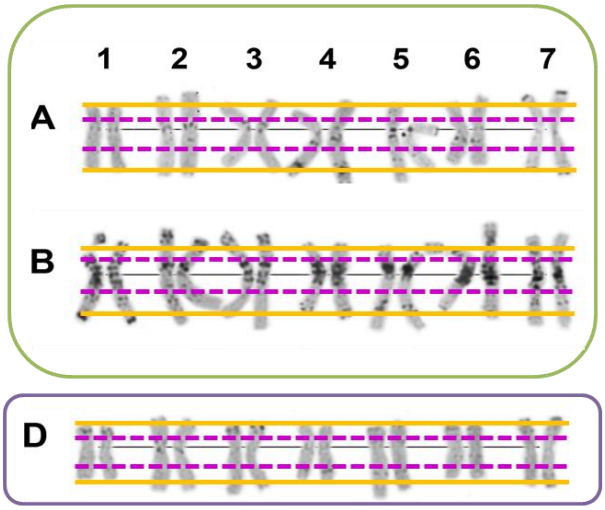
Ainsi, la technique de ChIP-seq semble très prometteuse pour l'identification et l'analyse des états chromatiniens chez les plantes. Cependant, Nakato et Shirahige présentaient en 2017 le ChIP-seq comme une technique puissante mais ne possédant pas encore de flux opérationnel (« work flow ») optimal en toutes circonstances. Une étude méthodologique de reproductibilité spécifique à la réponse de différents tissus aux différentes étapes du protocole serait nécessaire pour définir des recommandations d'usage afin de produire des données de ChIP-seq fiables chez les plantes. En particulier, je suggère que 1) l'hétérogénéité des résultats liée à l'hétérogénéité des tissus soit compensée par une augmentation du nombre de réplicas biologiques et 2) que la reproductibilité des différentes étapes du protocole soit estimée et mentionnée. Ces données permettraient d'augmenter la confiance dans les résultats obtenus dans différentes études voire leur comparaison. Ce prérequis est essentiel pour aller vers la caractérisation des spécificités de marquage épigénétiques développementaux versus signatures épigénétiques évolutives.

Conclusion sur l'ensemble des résultats

Caryotype du génome du blé tendre

Gènes dispensables; évolution rapide ; H3K27me3; sous-fonctionnalisation; expression tissus-spécifique ; biais d'expression importants

« Core genes » ; évolution lente ; H3K9ac; fonctions conservées; expression constitutive; peu de biais d'expression



➔Impact de l'histoire évolutive sur les expressions des sous-génomes

Détection d'une integration partielle des transcriptomes

Expression plus autonome ?

➔Biais d'expression entre sous- génomes homéologues stochastiques ; pas de dominance d'expression

➔Peu de pertes de gènes post-polyploïdisation; non biaisée en faveur de l'un des sous-génomes

➔Présence de nombreux gènes paralogues déjà diversifiés complexifiant les analyses

IV. Mise en perspective de l'ensemble des résultats : apport d'une évolution conceptuelle de l'étude des régulations épigénétiques

L'épigénétique constitue une discipline relativement récente dans la biologie puisque ses paradigmes fondateurs ont été édifiés aux alentours des années 40 du siècle dernier, notamment par le paléontologue, embryologiste, et généticien britannique Conrad Waddigton. Initialement énoncée comme l'ensemble des processus causaux qui participent à la construction progressive du phénotype à partir du génotype au cours du développement, la définition de l'épigénétique n'a cessé d'évoluer au cours du siècle dernier. En effet, selon que ce sont des chercheurs en biologie moléculaire, des embryologistes, des généticiens ou encore des théoriciens de l'évolution qui s'emparent du terme, la définition donnée à l'épigénétique et les questions posées différent sensiblement (tableau 11).

Tableau 11. Les différentes conceptions de l'épigénétique, leurs définitions, leurs champs de recherche et le problème qu'elles visent à résoudre. Nicoglou et Merlin 2017.

Conception	Définition	Champs de recherche	Problème
W-épi (l'épigénétique selon Waddington) Années 1930-40	Les mécanismes causaux impliqués dans le développement par lesquels les gènes produisent des effets phénotypiques	• Génétique classique et embryologie expérimentales • Biologie du développement	Développement (au niveau de l'organisme)
N-épi (l'épigénétique selon Nanney) Années 1950-60	Systèmes intégratifs auxiliaires régulant l'expression des potentialités génétiques	• Génétique chimique (moléculaire) et biologie du développement	Développement (au niveau de la cellule)
RH-épi (l'épigénétique selon Riggs et Holliday) Années 1970 à 1990-2000 &	Les changements héritables, par mitose et/ou méiose, de la fonction des gènes qui ne peuvent pas être expliqués par des changements de la séquence d'ADN	• Génétique et épigénétique moléculaire	Développement (au niveau moléculaire)
M-épi (l'épigénétique moléculaire) Années 2000-2010	Toute modification de la chromatine ayant un impact sur l'expression des gènes, que cette modification soit héritable ou pas		
ED-épi (l'épigénétique selon l'évo-dévo) Années 1990-2010	Les mécanismes développementaux (au dessus du niveau de la séquence d'ADN) qui sont à l'origine du phénotype et de ses modifications au cours de l'évolution	• Génétique du développement • Biologie évolutive du développement (évo-dévo) • Biologie des systèmes	L'origine de la variation phénotypique et l'interaction entre développement et évolution
ES-épi (l'épigénétique selon la Synthèse Etendue) Années 2000-2010	Mélange de N-épi & ED-épi Focalisation sur l'hérédité épigénétique transgénérationnelle	• Biologie évolutive du développement (évo-dévo) • Biologie de l'évolution • Biologie des systèmes	L'origine de la variation phénotypique et l'évolution vers une synthèse évolutionnaire étendue

Comprendre l'évolution des génomes polyploïdes par le prisme des processus épigénétiques implique de prendre en compte les caractéristiques du concept et de mettre en place des expériences robustes pour appréhender leur dynamique et leur caractère transitoire mais héritable. Ces processus sont à la base des changements d'expression du génome au cours du développement mais aussi en réponse à des stress. Les gènes d'un génome donné vont s'exprimer selon les programmes de modifications des marques

épigénétiques enregistrés qui vont s'opérer au cours du développement pour la différenciation des cellules, à l'image d'un orgue de Barbarie qui exécute son programme musical. Or, le déroulement des modifications de la chimie de la molécule d'ADN peut être influencé et modifié par les variations des conditions environnementales dans lesquelles il s'exécute. Permettant dans une certaine mesure, une adaptation en temps réel du déroulement du développement, cela implique également que ces modifications du déroulement du développement en temps réel soient transmises aux générations futures afin qu'elles puissent être potentiellement sélectionnées (si avantageuses) au cours de l'évolution. Certaines études montrent l'héritabilité de certaines marques épigénétiques sur plusieurs générations (Heard et Martienssen 2014, Perez et Lehner 2019) mais la question de l'emprise de la sélection naturelle sur ces processus très dynamiques et transitoires reste encore à explorer. L'analyse de l'évolution des gènes dans une perspective épigénétique nécessite ainsi de prendre en compte le cœur de la définition de l'épigénétique à savoir la modification de l'expression des gènes au cours du développement et en réponse à des stress.

Se poser la question de comment peut évoluer la redondance génétique d'un génome polyploïde peut alors revenir à se demander :

- si le déroulé des programmes épigénétiques des différents sous-génomes homéologues est synchrone
- si flexibilité de la synchronicité de l'expression des gènes est tolérée grâce à la redondance génétique
- si l'évolution de la régulation épigénétique d'un des *loci* au cours du développement est permise grâce à cette redondance et si elle peut entraîner l'apparition d'une variation du programme épigénétique

Ainsi, selon mon expérience de thèse, l'évolution épigénétique des gènes ne doit pas être étudiée en dehors des processus développementaux et des réponses aux stress environnementaux à travers lesquels se déroule l'ontogénèse d'un individu. D'où la nécessité de rendre très robuste les techniques de détection des signatures épigénétiques. Pour aller plus loin dans cette idée concernant l'utilisation du concept d'épigénétique dans les études d'évolution des espèces, je propose une réflexion conceptuelle, fondée sur le document produit par le CNRS « Epigénétique, Ecologie et évolution » (2018) et sur la publication de Klironomos *et al.* 2013, publications qui me semblent à l'avant-garde des futures études dans ce domaine.

1. Epigénétique, développement et évolution

À partir des années 1990, le courant disciplinaire de l'évo-dévo (l'évolution comprise sous le prisme la biologie du développement) a cherché à rendre compte de l'importance du processus développemental pour expliquer l'évolution (Denis Duboule 2018, *Le génome et ses embryons*). Selon ce courant, l'ontogénèse d'un organisme résume les étapes évolutives de l'espèce. Le développement de ce courant a permis l'émergence d'une synthèse étendue de la théorie de l'évolution qui a la particularité de prendre en

compte des variations non-génétiques et non-mendeliennes des caractères qui peuvent se produire au cours du développement et qui seront transmises à la descendance. Ainsi, si l'on tient compte des principes de l'évo-dévo, la décomposition des causes de la variance phénotypique au sein d'une population d'individus d'une même espèce permettrait en théorie d'identifier les différentes sources de variabilité héritable et non-héritable des caractères au cours du développement. Cela permettrait aussi d'aller vers une quantification de la part de chaque processus (génétiques, épigénétiques, lien entre les deux) dans la plasticité phénotypique héritée et soumise à pression de sélection.

Jusque récemment, les composantes génétiques et épigénétiques de l'hérédité étaient souvent perçues comme opposées et n'expliquant pas les mêmes phénomènes. Les premières étaient considérées comme base de l'évolution des espèces et les secondes comme potentiel adaptatif transitoire. Les sources de variabilité phénotypique étaient souvent résumées par l'équation **P = GxE** (ou P=phénotype, G= génotype et E= environnement). Prendre en compte les variations épigénétiques permettrait d'augmenter (au sens d'apporter un supplément) le concept P = GxE et de remplacer le G par **« système d'héritabilité »** : **P=SHxE** (Cosseau *et al.* 2017, cahier prospective CNRS). Ce système d'héritabilité comprendrait alors :

- le génotype : gènes, éléments de régulation en *cis*, environnement en TE entre autre
- l'épigénotype : les facteurs moléculaires autour de la séquence qui conditionnent l'accessibilité du génotype pour la machinerie de transcription et aux facteurs de transcription
- le compartiment cytoplasmique (petits ARN maternels par exemple)
- les microorganismes symbiotiques, parasites ou commensaux
- etc.

Selon cette proposition, l'idée est de produire un cadre conceptuel plus large pour comprendre l'évolution des espèces, permettant notamment d'intégrer les systèmes de mémorisation de l'environnement dans lequel se réalise le développement des individus d'une espèce ; développement qui peut être perturbé mais qui s'adapte grâce aux modifications épigénétiques induites.

2. Epigénétique, environnement, développement, évolution des espèces polyploïdes et amélioration variétale

Selon plusieurs chercheurs (Van de Peer, Alix, Freeling), la distribution spatio temporelle des espèces polyploïdes n'est pas aléatoire et se trouve corrélée à 1) des environnements extrêmes, 2) des crises environnementales affectant la biodiversité, 3) la colonisation de nouveaux biotopes. Selon Freeling, les individus polyploïdes constitueraient des pilliers permettant la survie d'une espèce. D'arpès ce chercheur et d'autres, l'occurrence de la polyploïdisation est plus fréquente lors d'évenements environnementaux extrêmes qui entraînent une production de gamètes non réduits plus importante (également démontré par Lokhande *et al.* 2003, Freeling *et al.* 2009 De Storme and Geelen 2013, Mirzaghaderi and Horandl 2016, Fawcett *et al.* 2017). La polyploïdisation entraînant la réunion de copies de gènes de la même fonction au sein d'un même génome, cette redondance génétique peut conduire à un relâchement de la préssion de

sélection sur certaines séquences et à des innovations évolutives plus rapides. Si les traces de cette évolution au cours des millions d'années sont conservées au sein de la séquence, il est possible que certaines innovations phénotypiques soient le fait de modifications épigénétiques entraînant un développement plus adapté aux environnements extrêmes dans les premières générations exposées et qui peuvent être héritées. Ceci peut être d'autant plus favorable à la survie d'une espèce polyploïde car selon Freeling toujours, ces espèces présentent une forte propension à la reproduction végétative assurant la survie de ces espèces dans des environnements extrêmes (Zhang *et al*. 2019. Ces deux processus (génétique et épigénétique) ne sont donc pas à considérer comme ayant des rôles distincts chez les organismes vivants mais simplement agissant à des échelles de temps différentes dans l'évolution des espèces.

Ces considérations conceptuelles sont pour moi un enjeu crucial au regard des questions posées et des solutions proposées dans le cadre de la recherche sur l'amélioration variétale. En effet, la prise en compte des caractéristiques du paradigme épigénétique telles que développées plus haut permettrait également de faire évoluer les pratiques d'amélioration variétale et de réassocier les cultures à leur environnement en utilisant le potentiel d'adaptabilité épigénétique des plantes.

3. Synthèse et conclusion

A partir de tout ce que j'ai pu découvrir au cours de mes travaux de recherche durant cette thèse, concernant les phénomènes de polyploïdisation et les processus épigénétiques, ont émergées au cours de mes réflexions des questions concernant la théorie de l'évolution des espèces. Je me suis grandement interrogée sur ce que pourraient expliquer ces processus au sein de cette théorie. Je présente un modèle synthétique dans la Figure 44 ci-après qui résume cette théorie de l'évolution augmentée par la prise en compte des processus épigénétiques. Je propose ainsi que ces derniers soient appréhendés comme des facteurs favorisant la diversification des processus développementaux au cours de l'ontogénèse des individus d'une espèce, la sélection naturelle s'opérerant à tout moment de la vie de l'individu selon le degré d'adpatation de ce dernier à son environnement. A la fin du développement, les gamètes peuvent présenter des variations épigénétiques qui correspondent à des modifications enregistrées lors de la méiose (et au cours de la vie du sporophyte) qui peuvent être transmises au gamétophyte chez les plantes.

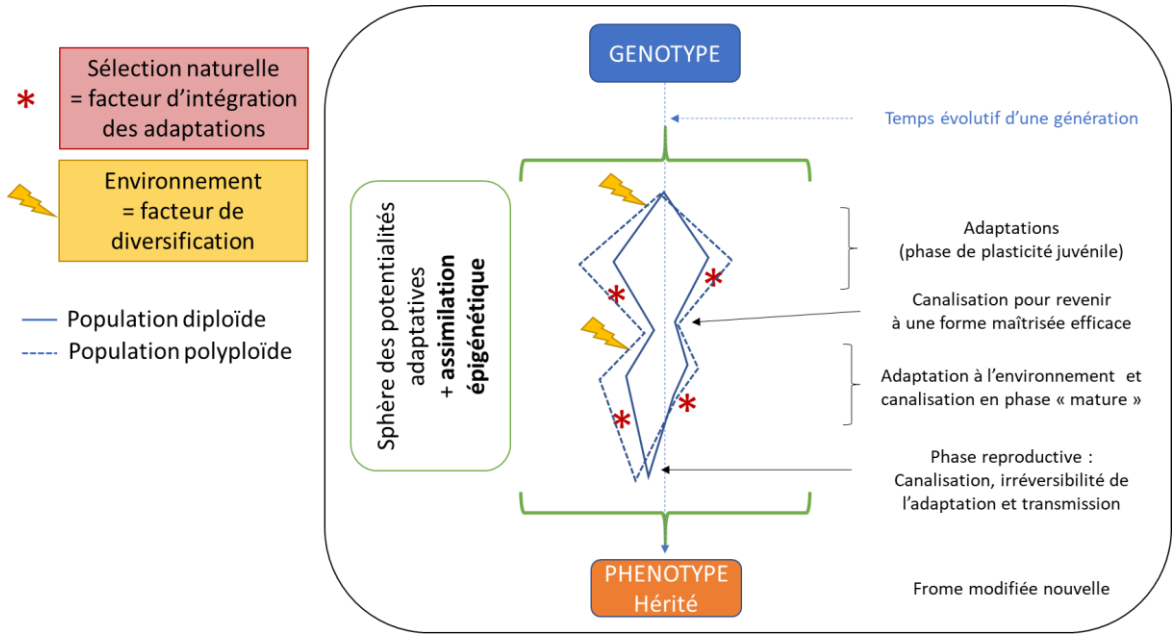


Figure 45. Schéma hypothétique de l'implication de la régulation épigénétique dans l'adaptabilité du processus de développement..

Ainsi, la cellule garante de la reproduction de l'espèce peut avoir enregistré des variations épigénétiques induites donnant un léger décalage phénotypique avec le phénotype modèle de l'espèce. Ces variations peuvent alors se transmettre à la génération suivante. Si l'on introduit ce schéma à travers le principe récursif du processus de polyploïdisation/diploïdisation, la possibilité d'obtenir des variations phénotypiques liées à des processus épigénétiques est démultipliée face aux variations environnementales chez des individus polyploïdes grâce à la redondance génétique et épigénétique. Ce schéma fait écho au sablier phylotypique de Raff (1996) qui propose deux types de contraintes évolutives lors de l'ontogénèse d'un individu :

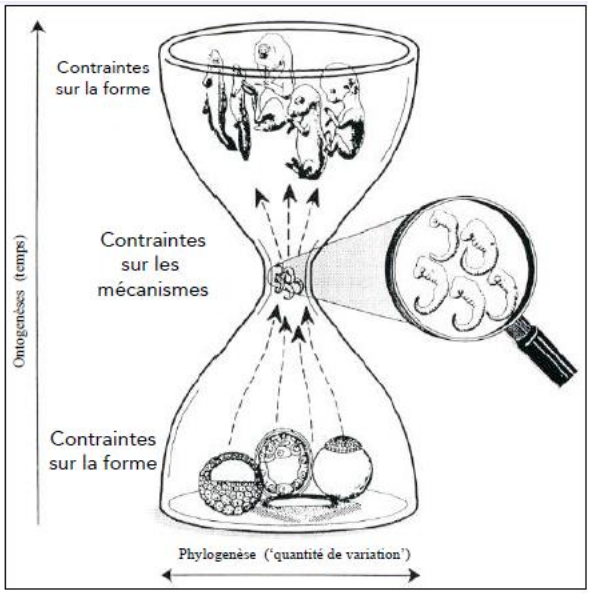


Figure 46. Le sablier phylotypique (Raff 1996, adapté par Duboule 2018)

contraintes sur la forme engeandrant des variations potentielles de phénotypes et des contraintes sur les mécanismes du développement contraignant et minimisant l'écart au phénotype « attendu » de l'espèce.

Ces considérations renvoient aux questions déjà posées par un certains nombre de chercheurs : les patrons de micro-évolution progressifs s'appliquant à l'échelle d'un individu ou d'une population mais surtout sur un pas de temps correspondant à une ou quelques générations, peuvent-ils avoir une incidence sur la macro-évolution des espèces à l'échelle des temps évolutifs ?

Une expérimentation sur une longue échelle de temps avec le suivi en parallèle des remaniements de séquences, des patrons de régulation épigénétique, des phénotypes des espèces diploides progénénitrices et

des descendants polyploïdes au cours de plusieurs générations et au sein d'une population évoluant dans un contexte naturel permettrait de décortiquer les variations phénotypiques observées et d'identifier la part de chacun des processus (systèmes d'héritabilité) dans l'établissement des phénotypes. Ce type d'expérience permettrait en outre de déterminer la part de chaque élément du système d'héritabilité dans l'établissement des innovations adaptatives et donc de l'évolution des espèces.

Après ces quelques années passées dans différents laboratoires de recherche ou en accompagnant des élèves dans leur compréhension du monde vivant, j'ai pu apprécier directement et plus indirectement comment les formes vivantes sont le fruit de différents cycles se répétant indéfiniement mais avec des potentialités d'évolution des formes de vies prenant des chemins très diversifiés. Etant très sensible aux liens entre arts et sciences qui ne sont que deux moyens différents d'explorer une même réalité, j'ai rapproché ma vision de l'évolution, qui consiste en la récursivité des processus maléables du vivant, aux travaux artistiques de **Maurits Cornelis Escher** qui a exploré les possibilités de création de formes nouvelles en modifiant les règles de la perspective (notamment par l'utilisation des **fractales** et de la perspective cylindrique). *Cette vision artistique traduit pour moi les possibilités infinies de l'évolution de la matière vivante.*

M. C. Escher, *Mosaic II, **1957***

REFERENCES BIBLIOGRAPHIQUE

Chapitre I : Synthèse bibliographique

Acharya, D., & Ghosh, T. C. (2016). Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics*, *17*(1), 71. https://doi.org/10.1186/s12864-016-2392-0

Alix, K., Gérard, P. R., Schwarzacher, T., & Heslop-Harrison, J. (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of botany*, 120(2), 183–194. https://doi.org/10.1093/aob/mcx079

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., & Quesneville, H. (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Research*, *39*(16), 6919-6931. https://doi.org/10.1093/nar/gkr324

Ainouche, M. L., Baumel, A., Salmon, A., & Yannic, G. (2004). Hybridization, polyploidy and speciation in Spartina (Poaceae). *New Phytologist*, *161*(1), 165-172. https://doi.org/10.1046/j.1469-8137.2003.00926.x

Akhunova, A. R., Matniyazov, R. T., Liang, H., & Akhunov, E. D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics*, *11*, 505. https://doi.org/10.1186/1471-2164-11-505

Albertin, W., & Marullo, P. (2012). Polyploidy in fungi : Evolution after whole-genome duplication. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), 2497-2509. https://doi.org/10.1098/rspb.2012.0434

Alix, K., Joets, J., Ryder, C. D., Moore, J., Barker, G. C., Bailey, J. P., King, G. J., & Heslop-Harrison, J. S. (Pat). (2008). The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation. *The Plant Journal*, *56*(6), 1030-1044. https://doi.org/10.1111/j.1365-313X.2008.03660.x

Altenhoff, A. M., Glover, N. M., & Dessimoz, C. (2019). Inferring Orthology and Paralogy. *Methods in Molecular Biology (Clifton, N.J.)*, *1910*, 149-175. https://doi.org/10.1007/978-1-4939-9074-0_5

Arrigo, N., & Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology*, *15*(2), 140-146. https://doi.org/10.1016/j.pbi.2012.03.010

Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F., & Yant, L. (2018). The "Polyploid Hop" : Shifting Challenges and Opportunities Over the Evolutionary Lifespan of Genome Duplications. *Frontiers in Ecology and Evolution*, *6*. https://doi.org/10.3389/fevo.2018.00117

Baker, K., Dhillon, T., Colas, I., Cook, N., Milne, I., Milne, L., Bayer, M., & Flavell, A. J. (2015). Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance. *The Plant Journal*, *84*(1), 111-124. https://doi.org/10.1111/tpj.12963

Balfourier, F., Bouchet, S., Robert, S., Oliveira, R. D., Rimbert, H., Kitt, J., Choulet, F., Consortium, I. W. G. S., Consortium, B., & Paux, E. (2019). Worldwide phylogeography and history of wheat genetic diversity. *Science Advances*, *5*(5), eaav0536. https://doi.org/10.1126/sciadv.aav0536

Bardil, A., de Almeida, J. D., Combes, M. C., Lashermes, P., & Bertrand, B. (2011). Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. *The New Phytologist*, *192*(3), 760-774. https://doi.org/10.1111/j.1469-8137.2011.03833.x

Barth, T. K., & Imhof, A. (2010). Fast signals and slow marks : The dynamics of histone modifications. *Trends in Biochemical Sciences*, *35*(11), 618-626. https://doi.org/10.1016/j.tibs.2010.05.006

Baumel, A., Ainouche, M. L., & Levasseur, J. E. (2001). Molecular investigations in populations of Spartina anglica C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Molecular*

Ecology, *10*(7), 1689-1701. https://doi.org/10.1046/j.1365-294x.2001.01299.x

Behling, A. H., Shepherd, L. D., & Cox, M. P. (2020). The importance and prevalence of allopolyploidy in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand*, *50*(2), 189-210. https://doi.org/10.1080/03036758.2019.1676797

Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M., & Jacobsen, S. E. (2008). Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE*, *3*(9). https://doi.org/10.1371/journal.pone.0003156

Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis : Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 14746-14753. https://doi.org/10.1073/pnas.1207726109

Bird, K. A., VanBuren, R., Puzey, J. R., & Edger, P. P. (2018). The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist*, *220*(1), 87-93. https://doi.org/10.1111/nph.15256

Blaine Marchant, D., Soltis, D. E., & Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *The New Phytologist*, *212*(3), 708-718. https://doi.org/10.1111/nph.14069

Bottani, S., Zabet, N. R., Wendel, J. F., & Veitia, R. A. (2018). Gene Expression Dominance in Allopolyploids : Hypotheses and Models. *Trends in Plant Science*, *23*(5), 393-402. https://doi.org/10.1016/j.tplants.2018.01.002

Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A.-C., & Elven, R. (2004). Polyploidy in arctic plants. *Biological Journal of the Linnean Society*, *82*(4), 521-536. https://doi.org/10.1111/j.1095-8312.2004.00337.x

Bruggmann, R., Bharti, A. K., Gundlach, H., Lai, J., Young, S., Pontaroli, A. C., Wei, F., Haberer, G., Fuks, G., Du, C., Raymond, C., Estep, M. C., Liu, R., Bennetzen, J. L., Chan, A. P., Rabinowicz, P. D., Quackenbush, J., Barbazuk, W. B., Wing, R. A., … Messing, J. (2006). Uneven chromosome contraction and expansion in the maize genome. *Genome Research*, *16*(10), 1241-1251. https://doi.org/10.1101/gr.5338906

Buzas, D. M. (2017). Capturing Environmental Plant Memories in DNA, with a Little Help from Chromatin. *Plant and Cell Physiology*, *58*(8), 1302-1312. https://doi.org/10.1093/pcp/pcx092

Cai, C., Wang, X., Liu, B., Wu, J., Liang, J., Cui, Y., Cheng, F., & Wang, X. (2017). Brassica rapa Genome 2.0 : A Reference Upgrade through Sequence Re-assembly and Gene Re-annotation. *Molecular Plant*, *10*(4), 649-651. https://doi.org/10.1016/j.molp.2016.11.008

Chagué, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., & Chalhoub, B. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. *The New Phytologist*, *187*(4), 1181-1194. https://doi.org/10.1111/j.1469-8137.2010.03339.x

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Da Silva, C., Just, J., Falentin, C., Koh, C. S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., … Wincker, P. (2014a). Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science (New York, N.Y.)*, *345*(6199), 950-953. https://doi.org/10.1126/science.1253435

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Silva, C. D., Just, J., Falentin, C., Koh, C. S., Clainche, I. L., Bernard, M., Bento, P., Noel, B., … Wincker, P. (2014b). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, *345*(6199), 950-953. https://doi.org/10.1126/science.1253435

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Silva, C. D., Just, J., Falentin, C., Koh, C. S., Clainche, I. L., Bernard, M., Bento, P., Noel, B., … Wincker, P. (2014c). Early allopolyploid

evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, *345*(6199), 950-953. https://doi.org/10.1126/science.1253435

Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B., & Salt, D. E. (2013). Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science (New York, N.Y.)*, *341*(6146), 658-659. https://doi.org/10.1126/science.1240561

Chelaifa, H., Chagué, V., Chalabi, S., Mestiri, I., Arnaud, D., Deffains, D., Lu, Y., Belcram, H., Huteau, V., Chiquet, J., Coriton, O., Just, J., Jahier, J., & Chalhoub, B. (2013). Prevalence of gene expression additivity in genetically stable wheat allohexaploids. *New Phytologist*, *197*(3), 730-736. https://doi.org/10.1111/nph.12108

Chelaifa, H., Monnier, A., & Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina x townsendii and Spartina anglica (Poaceae). *The New Phytologist*, *186*(1), 161-174. https://doi.org/10.1111/j.1469-8137.2010.03179.x

Chen, Z. J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual review of plant biology*, *58*, 377-406. https://doi.org/10.1146/annurev.arplant.58.032806.103835

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, *4*(5), 258-268. https://doi.org/10.1038/s41477-018-0136-7

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*(5), e36442. https://doi.org/10.1371/journal.pone.0036442

Cheng, F., Wu, J., & Wang, X. (2014). Genome triplication drove the diversification of Brassica plants. *Horticulture Research*, *1*(1), 1-8. https://doi.org/10.1038/hortres.2014.24

Chester, M., Gallagher, J. P., Symonds, V. V., Cruz da Silva, A. V., Mavrodiev, E. V., Leitch, A. R., Soltis, P. S., & Soltis, D. E. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America*, *109*(4), 1176-1181. https://doi.org/10.1073/pnas.1112041109

Chittock, E. C., Latwiel, S., Miller, T. C. R., & Müller, C. W. (2017). Molecular architecture of polycomb repressive complexes. *Biochemical Society Transactions*, *45*(1), 193-205. https://doi.org/10.1042/BST20160173

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., Gouis, J. L., Balfourier, F., Alaux, M., Jamilloux, V., Poulain, J., Durand, C., … Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science*, *345*(6194). https://doi.org/10.1126/science.1249721

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Paslier, M.-C. L., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., … Feuillet, C. (2010). Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. *The Plant Cell*, *22*(6), 1686-1701. https://doi.org/10.1105/tpc.110.074187

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews. Genetics*, *6*(11), 836-846. https://doi.org/10.1038/nrg1711

Combes Gavalda, M.-C. (2015). *Polyploïdie et adaptation des plantes : Caractérisation et variation de l'expression des gènes homoélogues chez le caféier Coffea arabica* [These de doctorat, Montpellier]. http://www.theses.fr/2015MONTS115

Conant, G. C., Birchler, J. A., & Pires, J. C. (2014). Dosage, duplication, and diploidization : Clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology*, *19*, 91-98. https://doi.org/10.1016/j.pbi.2014.05.008

Consortium (IWGSC), T. I. W. G. S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, *345*(6194). https://doi.org/10.1126/science.1251788

Consortium (IWGSC), T. I. W. G. S., Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403). https://doi.org/10.1126/science.aar7191

Crow, K. D., & Wagner, G. P. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, *23*(5), 887-892. https://doi.org/10.1093/molbev/msj083

De Storme, N., & Mason, A. (2014). Plant speciation through chromosome instability and ploidy change : Cellular mechanisms, molecular factors and evolutionary relevance. *Current Plant Biology*, *1*, 10-33. https://doi.org/10.1016/j.cpb.2014.09.002

Deal, R. B., & Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nature Protocols*, *6*(1), 56-68. https://doi.org/10.1038/nprot.2010.175

Derkacheva, M., & Hennig, L. (2014). Variations on a theme : Polycomb group proteins in plants. *Journal of Experimental Botany*, *65*(10), 2769-2784. https://doi.org/10.1093/jxb/ert410

Diderot, D., & Briasson. (1765). *Encyclopédie : Ou dictionnaire raisonné des sciences, des arts et des métiers*.

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., & Wendel, J. F. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics*, *42*, 443-461. https://doi.org/10.1146/annurev.genet.42.110807.091524

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., Heckel, D. G., Der, J. P., Wafula, E. K., Tang, M., Hofberger, J. A., Smithson, A., Hall, J. C., Blanchette, M., Bureau, T. E., Wright, S. I., dePamphilis, C. W., Schranz, M. E., Barker, M. S., … Wheat, C. W. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences*, *112*(27), 8362-8366. https://doi.org/10.1073/pnas.1503926112

Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., Bowman, M. J., Childs, K. L., Washburn, J. D., Schmitz, R. J., Smith, G. D., Pires, J. C., & Puzey, J. R. (2017). Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower[OPEN]. *The Plant Cell*, *29*(9), 2150-2167. https://doi.org/10.1105/tpc.17.00010

Eichten, S. R., Swanson-Wagner, R. A., Schnable, J. C., Waters, A. J., Hermanson, P. J., Liu, S., Yeh, C.-T., Jia, Y., Gendler, K., Freeling, M., Schnable, P. S., Vaughn, M. W., & Springer, N. M. (2011). Heritable epigenetic variation among maize inbreds. *PLoS Genetics*, *7*(11), e1002372. https://doi.org/10.1371/journal.pgen.1002372

ENQ-CER-repvar-A19.pdf. (s. d.). Consulté 19 août 2020, à l'adresse https://www.franceagrimer.fr/fam/content/download/61615/document/ENQ-CER-repvar-A19.pdf?version=2

F, B., & Jd, T. (1995). Gametes with the somatic chromosome number : Mechanisms of their formation and role in the evolution of autopolyploid plants. *The New Phytologist*, *129*(1), 1-22. https://doi.org/10.1111/j.1469-8137.1995.tb03005.x

Feldman, M., & Levy, A. A. (2015). Origin and Evolution of Wheat and Related Triticeae Species. In M. Molnár-Láng, C. Ceoloni, & J. Doležel (Éds.), *Alien Introgression in Wheat : Cytogenetics, Molecular Biology, and Genomics* (p. 21-76). Springer International Publishing. https://doi.org/10.1007/978-3-319-23494-6_2

Felsenfeld, G. (2014). A Brief History of Epigenetics. *Cold Spring Harbor Perspectives in Biology*, *6*(1). https://doi.org/10.1101/cshperspect.a018200

Flagel, L. E., & Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *The New Phytologist*, *186*(1), 184-193. https://doi.org/10.1111/j.1469-8137.2009.03107.x

Food and Agriculture Organization of the United Nations, & Trade and Markets Division. (2018). *Food outlook : Biannual report on global food markets, November 2018.*

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., & Mummenhoff, K. (2011). Cabbage family affairs : The evolutionary history of Brassicaceae. *Trends in Plant Science*, *16*(2), 108-116. https://doi.org/10.1016/j.tplants.2010.11.005

Freeling, M., Scanlon, M. J., & Fowler, J. E. (2015). Fractionation and subfunctionalization following genome duplications : Mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development*, *35*, 110-118. https://doi.org/10.1016/j.gde.2015.11.002

Future Protein Supply and Demand : Strategies and Factors Influencing a Sustainable Equilibrium. (s. d.). Consulté 14 juillet 2020, à l'adresse https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5532560/

Gaeta, R. T., & Pires, J. C. (2010). Homoeologous recombination in allopolyploids : The polyploid ratchet. *New Phytologist*, *186*(1), 18-28. https://doi.org/10.1111/j.1469-8137.2009.03089.x

Gardiner, L.-J., Quinton-Tulloch, M., Olohan, L., Price, J., Hall, N., & Hall, A. (2015). A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology*, *16*(1), 273. https://doi.org/10.1186/s13059-015-0838-3

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., & Freeling, M. (2014). Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution*, *31*(2), 448-454. https://doi.org/10.1093/molbev/mst230

Gates, L. A., Shi, J., Rohira, A. D., Feng, Q., Zhu, B., Bedford, M. T., Sagum, C. A., Jung, S. Y., Qin, J., Tsai, M.-J., Tsai, S. Y., Li, W., Foulds, C. E., & O'Malley, B. W. (2017). Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *The Journal of Biological Chemistry*, *292*(35), 14456-14472. https://doi.org/10.1074/jbc.M117.802074

Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization – ScienceOpen. (s. d.). Consulté 17 août 2020, à l'adresse https://www.scienceopen.com/document?vid=8f124b52-615d-401b-b36d-afdc4c0ff2fa

Genome-specific differential gene expressions in resynthesized Brassica allotetraploids from pair-wise crosses of three cultivated diploids revealed by RNA-seq. (s. d.). Consulté 23 juillet 2020, à l'adresse https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631939/

Gent, J. I., Madzima, T. F., Bader, R., Kent, M. R., Zhang, X., Stam, M., McGinnis, K. M., & Dawe, R. K. (2014). Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation. *The Plant Cell*, *26*(12), 4903-4917. https://doi.org/10.1105/tpc.114.130427

Glover, N. M., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Feuillet, C., & Choulet, F. (2015). Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. *Genome Biology*, *16*(1), 188. https://doi.org/10.1186/s13059-015-0754-6

Glover, N. M., Redestig, H., & Dessimoz, C. (2016). Homoeologs : What Are They and How Do We Infer Them? *Trends in Plant Science*, *21*(7), 609-621. https://doi.org/10.1016/j.tplants.2016.02.005

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., & Wendel, J. F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids.

New Phytologist, *196*(4), 966-971. https://doi.org/10.1111/j.1469-8137.2012.04365.x

Grover, Corrinne E., Yu, Y., Wing, R. A., Paterson, A. H., & Wendel, J. F. (2008). A phylogenetic analysis of indel dynamics in the cotton genus. *Molecular Biology and Evolution*, *25*(7), 1415-1428. https://doi.org/10.1093/molbev/msn085

Guo, H., Jiao, Y., Tan, X., Wang, X., Huang, X., Jin, H., & Paterson, A. H. (2019). Gene duplication and genetic innovation in cereal genomes. *Genome Research*, *29*(2), 261-269. https://doi.org/10.1101/gr.237511.118

Hao, M., Li, A., Shi, T., Luo, J., Zhang, L., Zhang, X., Ning, S., Yuan, Z., Zeng, D., Kong, X., Li, X., Zheng, H., Lan, X., Zhang, H., Zheng, Y., Mao, L., & Liu, D. (2017). The abundance of homoeologue transcripts is disrupted by hybridization and is partially restored by genome doubling in synthetic hexaploid wheat. *BMC Genomics*, *18*. https://doi.org/10.1186/s12864-017-3558-0

Harper, A. L., Trick, M., He, Z., Clissold, L., Fellgett, A., Griffiths, S., & Bancroft, I. (2016). Genome distribution of differential homoeologue contributions to leaf gene expression in bread wheat. *Plant Biotechnology Journal*, *14*(5), 1207-1214. https://doi.org/10.1111/pbi.12486

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., Knox, R., Cuthbert, R., Pozniak, C., Akhunova, A., Morrell, P. L., Davies, J. P., Webb, S. R., Spangenberg, G., Hayes, B., … Akhunov, E. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nature Genetics*, *51*(5), 896-904. https://doi.org/10.1038/s41588-019-0382-2

He, F., Zhang, X., Hu, J., Turck, F., Dong, X., Goebel, U., Borevitz, J., & de Meaux, J. (2012). Genome-wide analysis of cis-regulatory divergence between species in the Arabidopsis genus. *Molecular Biology and Evolution*, *29*(11), 3385-3395. https://doi.org/10.1093/molbev/mss146

Henchion, M., Hayes, M., Mullen, A. M., Fenelon, M., & Tiwari, B. (2017). Future Protein Supply and Demand : Strategies and Factors Influencing a Sustainable Equilibrium. *Foods*, *6*(7). https://doi.org/10.3390/foods6070053

Henderson, I. R., & Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, *447*(7143), 418-424. https://doi.org/10.1038/nature05917

Hodkinson, T. R. (2018). Evolution and Taxonomy of the Grasses (Poaceae) : A Model Family for the Study of Species-Rich Groups. In *Annual Plant Reviews online* (p. 255-294). American Cancer Society. https://doi.org/10.1002/9781119312994.apr0622

Hu, Y., Lu, Y., Zhao, Y., & Zhou, D.-X. (2019). Histone Acetylation Dynamics Integrates Metabolic Activity to Regulate Plant Response to Stress. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01236

Hu, Z., Song, N., Xing, J., Chen, Y., Han, Z., Yao, Y., Peng, H., Ni, Z., & Sun, Q. (2013). Overexpression of Three TaEXPA1 Homoeologous Genes with Distinct Expression Divergence in Hexaploid Wheat Exhibit Functional Retention in Arabidopsis. *PLOS ONE*, *8*(5), e63667. https://doi.org/10.1371/journal.pone.0063667

Hughes, T. E., Langdale, J. A., & Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Research*, *24*(8), 1348-1355. https://doi.org/10.1101/gr.172684.114

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., Pires, J. C., Chalhoub, B., King, G. J., Snowdon, R., Batley, J., & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnology Journal*, *16*(7), 1265-1274. https://doi.org/10.1111/pbi.12867

Jackson, S. A. (2017). Epigenomics : Dissecting hybridization and polyploidization. *Genome Biology*, *18*. https://doi.org/10.1186/s13059-017-1254-7

Jiang, J., Shao, Y., Du, K., Ran, L., Fang, X., & Wang, Y. (2013). Use of digital gene expression to discriminate gene expression differences in early generations of resynthesized Brassica napus and its diploid progenitors. *BMC Genomics*, *14*, 72. https://doi.org/10.1186/1471-2164-14-72

Jiang, W., Liu, Y., Xia, E., & Gao, L. (2013). Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. *Plant Physiology*, *161*(4), 1844-1861. https://doi.org/10.1104/pp.112.200147

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., & dePamphilis, C. W. (2011a). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97-100. https://doi.org/10.1038/nature09916

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., & dePamphilis, C. W. (2011b). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97-100. https://doi.org/10.1038/nature09916

Johannes, F., Colot, V., & Jansen, R. C. (2008a). Epigenome dynamics : A quantitative genetics perspective. *Nature Reviews Genetics*, *9*(11), 883-890. https://doi.org/10.1038/nrg2467

Johannes, F., Colot, V., & Jansen, R. C. (2008b). Epigenome dynamics : A quantitative genetics perspective. *Nature Reviews Genetics*, *9*(11), 883-890. https://doi.org/10.1038/nrg2467

Kellogg, E. A. (2001). Evolutionary History of the Grasses. *Plant Physiology*, *125*(3), 1198-1205. https://doi.org/10.1104/pp.125.3.1198

Kilian, B., Martin, W., & Salamini, F. (2010). Genetic Diversity, Evolution and Domestication of Wheat and Barley in the Fertile Crescent. In M. Glaubrecht (Éd.), *Evolution in Action : Case studies in Adaptive Radiation, Speciation and the Origin of Biodiversity* (p. 137-166). Springer. https://doi.org/10.1007/978-3-642-12425-9_8

Kim, K. D., El Baidouri, M., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grimwood, J., & Jackson, S. A. (2015). A Comparative Epigenomic Analysis of Polyploidy-Derived Genes in Soybean and Common Bean1[OPEN]. *Plant Physiology*, *168*(4), 1433-1447. https://doi.org/10.1104/pp.15.00408

Kooke, R., Johannes, F., Wardenaar, R., Becker, F., Etcheverry, M., Colot, V., Vreugdenhil, D., & Keurentjes, J. J. B. (2015). Epigenetic basis of morphological variation and phenotypic plasticity in Arabidopsis thaliana. *The Plant Cell*, *27*(2), 337-348. https://doi.org/10.1105/tpc.114.133025

Lashermes, P., Combes, M.-C., Hueber, Y., Severac, D., & Dereeper, A. (2014). Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. *The Plant Journal*, *78*(4), 674-685. https://doi.org/10.1111/tpj.12505

Lavarone, E., Barbieri, C. M., & Pasini, D. (2019). Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nature Communications*, *10*(1), 1679. https://doi.org/10.1038/s41467-019-09624-w

Leal-Bertioli, S. C. M., Moretzsohn, M. C., Santos, S. P., Brasileiro, A. C. M., Guimarães, P. M., Bertioli, D. J., & Araujo, A. C. G. (2017). Phenotypic effects of allotetraploidization of wild Arachis and their implications for peanut domestication. *American Journal of Botany*, *104*(3), 379-388. https://doi.org/10.3732/ajb.1600402

Lehti-Shiu, M. D., Panchy, N., Wang, P., Uygun, S., & Shiu, S.-H. (2017). Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms*, *1860*(1), 3-20. https://doi.org/10.1016/j.bbagrm.2016.08.005

Leitch, A. R., & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science (New York, N.Y.)*, *320*(5875), 481-483. https://doi.org/10.1126/science.1153585

Leroy, P., Guilhot, N., Sakai, H., Bernard, A., Choulet, F., Theil, S., Reboux, S., Amano, N., Flutre, T., Pelegrin, C., Ohyanagi, H., Seidel, M., Giacomoni, F., Reichstadt, M., Alaux, M., Gicquello, E., Legeai, F., Cerutti, L., Numa, H., … Feuillet, C. (2012). TriAnnot : A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes. *Frontiers in Plant Science*, *3*, 5. https://doi.org/10.3389/fpls.2012.00005

Levin, D. A. (2019). Plant speciation in the age of climate change. *Annals of Botany*, *124*(5), 769-775. https://doi.org/10.1093/aob/mcz108

Li, J., Wan, H.-S., & Yang, W.-Y. (2014). Synthetic hexaploid wheat enhances variation and adaptive evolution of bread wheat in breeding processes. *Journal of Systematics and Evolution*, *52*(6), 735-742. https://doi.org/10.1111/jse.12110

Li, Q., Gent, J. I., Zynda, G., Song, J., Makarevitch, I., Hirsch, C. D., Hirsch, C. N., Dawe, R. K., Madzima, T. F., McGinnis, K. M., Lisch, D., Schmitz, R. J., Vaughn, M. W., & Springer, N. M. (2015). RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(47), 14728-14733. https://doi.org/10.1073/pnas.1514680112

Li, Zheng, Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). Early genome duplications in conifers and other seed plants. *Science Advances*, *1*(10), e1501084. https://doi.org/10.1126/sciadv.1501084

Li, Zheng, Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., & Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(18), 4713-4718. https://doi.org/10.1073/pnas.1710791115

Li, Zijuan, Wang, M., Lin, K., Xie, Y., Guo, J., Ye, L., Zhuang, Y., Teng, W., Ran, X., Tong, Y., Xue, Y., Zhang, W., & Zhang, Y. (2019). The bread wheat epigenomic map reveals distinct chromatin architectural and evolutionary features of functional genetic elements. *Genome Biology*, *20*(1), 139. https://doi.org/10.1186/s13059-019-1746-8

Liang, Z., & Schnable, J. C. (2018). Functional Divergence between Subgenomes and Gene Pairs after Whole Genome Duplications. *Molecular Plant*, *11*(3), 388-397. https://doi.org/10.1016/j.molp.2017.12.010

Lim, K. Yoong, Kovarik, A., Matyasek, R., Chase, M. W., Clarkson, J. J., Grandbastien, M. A., & Leitch, A. R. (2007). Sequence of events leading to near-complete genome turnover in allopolyploid Nicotiana within five million years. *The New Phytologist*, *175*(4), 756-763. https://doi.org/10.1111/j.1469-8137.2007.02121.x

Lim, Kar Yoong, Matyasek, R., Kovarik, A., & Leitch, A. R. (2004). Genome evolution in allotetraploid Nicotiana : GENOME EVOLUTION IN ALLOTETRAPLOID NICOTIANA. *Biological Journal of the Linnean Society*, *82*(4), 599-606. https://doi.org/10.1111/j.1095-8312.2004.00344.x

Lloyd, A., Blary, A., Charif, D., Charpentier, C., Tran, J., Balzergue, S., Delannoy, E., Rigaill, G., & Jenczewski, E. (2018). Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. *The New Phytologist*, *217*(1), 367-377. https://doi.org/10.1111/nph.14836

Mable, B. K. (2004). 'Why polyploidy is rarer in animals than in plants' : Myths and mechanisms: MYTHS AND MECHANISMS. *Biological Journal of the Linnean Society*, *82*(4), 453-466. https://doi.org/10.1111/j.1095-8312.2004.00332.x

Mable, B. K., & Otto, S. P. (2001). Masking and purging mutations following EMS treatment in haploid, diploid and tetraploid yeast (Saccharomyces cerevisiae). *Genetical Research*, *77*(1), 9-26. https://doi.org/10.1017/s0016672300004821

Mable, Barbara K., Brysting, A. K., Jørgensen, M. H., Carbonell, A. K. Z., Kiefer, C., Ruiz-Duarte, P., Lagesen, K., & Koch, M. A. (2018). Adding Complexity to Complexity : Gene Family Evolution in Polyploids. *Frontiers in Ecology and Evolution*, *6*.

https://doi.org/10.3389/fevo.2018.00114
Madlung, A. (2013). Polyploidy and its effect on evolutionary success : Old questions revisited with new tools. *Heredity*, *110*(2), 99-104. https://doi.org/10.1038/hdy.2012.79
Makarevitch, I., Eichten, S. R., Briskine, R., Waters, A. J., Danilevskaya, O. N., Meeley, R. B., Myers, C. L., Vaughn, M. W., & Springer, N. M. (2013). Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27. *The Plant Cell*, *25*(3), 780-793. https://doi.org/10.1105/tpc.112.106427
Mandáková, T., & Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, *42*, 55-65. https://doi.org/10.1016/j.pbi.2018.03.001
Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., & Olsen, O.-A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science (New York, N.Y.)*, *345*(6194), 1250092. https://doi.org/10.1126/science.1250092
Markov, A. V., & Kaznacheev, I. S. (2016). Evolutionary consequences of polyploidy in prokaryotes and the origin of mitosis and meiosis. *Biology Direct*, *11*, 28. https://doi.org/10.1186/s13062-016-0131-8
Marmagne, A., Brabant, P., Thiellement, H., & Alix, K. (2010). Analysis of gene expression in resynthesized Brassica napus allotetraploids : Transcriptional changes do not explain differential protein regulation. *The New Phytologist*, *186*(1), 216-227. https://doi.org/10.1111/j.1469-8137.2009.03139.x
Matsuoka, Y. (2011). Evolution of Polyploid Triticum Wheats under Cultivation : The Role of Domestication, Natural Hybridization and Allopolyploid Speciation in their Diversification. *Plant and Cell Physiology*, *52*(5), 750-764. https://doi.org/10.1093/pcp/pcr018
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. https://doi.org/10.1126/science.1207205
McGinty, R. K., & Tan, S. (2016). Recognition of the nucleosome by chromatin factors and enzymes. *Current opinion in structural biology*, *37*, 54-61. https://doi.org/10.1016/j.sbi.2015.11.014
McIntyre, P. J. (2012). Polyploidy associated with altered and broader ecological niches in the Claytonia perfoliata (Portulacaceae) species complex. *American Journal of Botany*, *99*(4), 655-662. https://doi.org/10.3732/ajb.1100466
Meirmans, P. G., Liu, S., & van Tienderen, P. H. (2018). The Analysis of Polyploid Genetic Data. *Journal of Heredity*, *109*(3), 283-296. https://doi.org/10.1093/jhered/esy006
Middleton, C. P., Senerchia, N., Stein, N., Akhunov, E. D., Keller, B., Wicker, T., & Kilian, B. (2014). Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PloS one*, 9(3), e85761. https://doi.org/10.1371/journal.pone.0085761
Mirouze, M., & Vitte, C. (2014). Transposable elements, a treasure trove to decipher epigenetic variation : Insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany*, *65*(10), 2801-2812. https://doi.org/10.1093/jxb/eru120
Muller, H. J. (1925). Why Polyploidy is Rarer in Animals Than in Plants. *The American Naturalist*, *59*(663), 346-353. https://doi.org/10.1086/280047
Mutti, J. S., Bhullar, R. K., & Gill, K. S. (2017). Evolution of Gene Expression Balance Among Homeologs of Natural Polyploids. *G3: Genes|Genomes|Genetics*, *7*(4), 1225-1237. https://doi.org/10.1534/g3.116.038711
Osborn, T. C., Butrulle, D. V., Sharpe, A. G., Pickering, K. J., Parkin, I. A. P., Parker, J. S., & Lydiate, D. J. (2003). Detection and effects of a homeologous reciprocal transposition in

Brassica napus. *Genetics*, *165*(3), 1569-1577.
Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, *131*(3), 452-462. https://doi.org/10.1016/j.cell.2007.10.022
Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 2896-2901.
Pandit, M. K., Pocock, M. J. O., & Kunin, W. E. (2011). Ploidy influences rarity and invasiveness in plants. *Journal of Ecology*, *99*(5), 1108-1115. https://doi.org/10.1111/j.1365-2745.2011.01838.x
Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., & Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. *New Phytologist*, *184*(4), 1003-1015. https://doi.org/10.1111/j.1469-8137.2009.03029.x
Parks, M. B., Nakov, T., Ruck, E. C., Wickett, N. J., & Alverson, A. J. (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany*, *105*(3), 330-347. https://doi.org/10.1002/ajb2.1056
Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., Lagudah, E., Somers, D., Kilian, A., Alaux, M., Vautrin, S., Bergès, H., Eversole, K., Appels, R., Safar, J., … Feuillet, C. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. *Science (New York, N.Y.)*, *322*(5898), 101-104. https://doi.org/10.1126/science.1161847
Pelé, A., Falque, M., Trotoux, G., Eber, F., Nègre, S., Gilet, M., Huteau, V., Lodé, M., Jousseaume, T., Dechaumet, S., Morice, J., Poncet, C., Coriton, O., Martin, O. C., Rousseau-Gueutin, M., & Chèvre, A.-M. (2017). Amplifying recombination genome-wide and reshaping crossover landscapes in Brassicas. *PLOS Genetics*, *13*(5), e1006794. https://doi.org/10.1371/journal.pgen.1006794
Pelé, A., Rousseau-Gueutin, M., & Chèvre, A.-M. (2018). Speciation Success of Polyploid Plants Closely Relates to the Regulation of Meiotic Recombination. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00907
Peng, J. H., Sun, D., & Nevo, E. (2011). Domestication evolution, genetics and genomics in wheat. *Molecular Breeding*, *28*(3), 281. https://doi.org/10.1007/s11032-011-9608-4
Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, Q. T., Lim, K. Y., Kovarik, A., Leitch, A. R., Grandbastien, M.-A., & Mhiri, C. (2010). Mobilization of retrotransposons in synthetic allotetraploid tobacco. *The New Phytologist*, *186*(1), 135-147. https://doi.org/10.1111/j.1469-8137.2009.03140.x
Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., International Wheat Genome Sequencing Consortium, Mayer, K. F. X., & Olsen, O.-A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science (New York, N.Y.)*, *345*(6194), 1250091. https://doi.org/10.1126/science.1250091
Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F., Molnár-Láng, M., Lage, J., Kilian, B., Özkan, H., Waite, D., Dyer, S., Letellier, T., Alaux, M., Wheat and Barley Legacy for Breeding Improvement (WHEALBI) consortium, … Salse, J. (2019). Tracing the ancestry of modern bread wheats. *Nature Genetics*, *51*(5), 905-911. https://doi.org/10.1038/s41588-019-0393-z
Pophaly, S. D., & Tellier, A. (2015). Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. *Molecular Biology and Evolution*, *32*(12), 3226-3235. https://doi.org/10.1093/molbev/msv191
Prentis, P. J., Wilson, J. R. U., Dormontt, E. E., Richardson, D. M., & Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends in Plant Science*, *13*(6), 288-294. https://doi.org/10.1016/j.tplants.2008.03.004

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, *20*(1), 38. https://doi.org/10.1186/s13059-019-1650-2

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J., & Zhang, S. (2018). Different Modes of Gene Duplication Show Divergent Evolutionary Patterns and Contribute Differently to the Expansion of Gene Families Involved in Important Fruit Traits in Pear (Pyrus bretschneideri). *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00161

Raju, S. K. K. (2020). Gene Dosage Balance Immediately following Whole-Genome Duplication in Arabidopsis[OPEN]. *The Plant Cell*, *32*(5), 1344-1345. https://doi.org/10.1105/tpc.20.00205

Ramsey, J., & Ramsey, T. S. (2014). Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1648), 20130352. https://doi.org/10.1098/rstb.2013.0352

Ramsey, J., & Schemske, D. W. (1998). Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants. *Annual Review of Ecology and Systematics*, *29*(1), 467-501. https://doi.org/10.1146/annurev.ecolsys.29.1.467

Rapp, R. A., Udall, J. A., & Wendel, J. F. (2009). Genomic expression dominance in allopolyploids. *BMC Biology*, *7*, 18. https://doi.org/10.1186/1741-7007-7-18

Renny-Byfield, S., & Wendel, J. F. (2014). Doubling down on genomes : Polyploidy and crop plants. *American Journal of Botany*, *101*(10), 1711-1725. https://doi.org/10.3732/ajb.1400119

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., … Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal*, *30*(10), 1928-1938. https://doi.org/10.1038/emboj.2011.103

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., & Jackson, S. A. (2013a). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal*, *73*(1), 143-153. https://doi.org/10.1111/tpj.12026

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., & Jackson, S. A. (2013b). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal: For Cell and Molecular Biology*, *73*(1), 143-153. https://doi.org/10.1111/tpj.12026

Rousseau-Gueutin, M., Morice, J., Coriton, O., Huteau, V., Trotoux, G., Nègre, S., Falentin, C., Deniot, G., Gilet, M., Eber, F., Pelé, A., Vautrin, S., Fourment, J., Lodé, M., Bergès, H., & Chèvre, A.-M. (2016). The Impact of Open Pollination on the Structural Evolutionary Dynamics, Meiotic Behavior, and Fertility of Resynthesized Allotetraploid Brassica napus L. *G3: Genes|Genomes|Genetics*, *7*(2), 705-717. https://doi.org/10.1534/g3.116.036517

Salamini, F., Ozkan, H., Brandolini, A., Schäfer-Pregl, R., & Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nature Reviews. Genetics*, *3*(6), 429-441. https://doi.org/10.1038/nrg817

Salman-Minkov, A., Sabath, N., & Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, *2*(8), 1-4. https://doi.org/10.1038/nplants.2016.115

Sarilar, V., Palacios, P. M., Rousselet, A., Ridel, C., Falque, M., Eber, F., Chèvre, A.-M., Joets, J., Brabant, P., & Alix, K. (2013). Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in resynthesized Brassica napus allotetraploids. *The New Phytologist*, *198*(2), 593-604. https://doi.org/10.1111/nph.12156

Schmid, M., Evans, B. J., & Bogart, J. P. (2015). Polyploidy in Amphibia. *Cytogenetic and Genome Research*, *145*(3-4), 315-330. https://doi.org/10.1159/000431388

Schnable, J. C., Springer, N. M., & Freeling, M. (2011). Differentiation of the maize subgenomes

by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, *108*(10), 4069-4074. https://doi.org/10.1073/pnas.1101368108

Schneider, H., Liu, H.-M., Chang, Y.-F., Ohlsen, D., Perrie, L. R., Shepherd, L., Kessler, M., Karger, D. N., Hennequin, S., Marquardt, J., Russell, S., Ansell, S., Lu, N. T., Kamau, P., Lóriga, J., Regalado, L., Heinrichs, J., Ebihara, A., Smith, A. R., & Gibby, M. (2017). Neo- and Paleopolyploidy contribute to the species diversity of Asplenium—The most species-rich genus of ferns. *Journal of Systematics and Evolution*, *55*(4), 353-364. https://doi.org/10.1111/jse.12271

Schreiber, A. W., Hayden, M. J., Forrest, K. L., Kong, S. L., Langridge, P., & Baumann, U. (2012). Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat. *BMC Genomics*, *13*, 492. https://doi.org/10.1186/1471-2164-13-492

Schwartz, Y. B., & Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. *Nature Reviews Genetics*, *8*(1), 9-22. https://doi.org/10.1038/nrg1981

Secco, D., Wang, C., Shou, H., Schultz, M. D., Chiarenza, S., Nussaume, L., Ecker, J. R., Whelan, J., & Lister, R. (2015). Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *ELife*, *4*. https://doi.org/10.7554/eLife.09343

Senerchia, N., Wicker, T., Felber, F., & Parisod, C. (2013). Evolutionary Dynamics of Retrotransposons Assessed by High-Throughput Sequencing in Wild Relatives of Wheat. *Genome Biology and Evolution*, *5*(5), 1010-1020. https://doi.org/10.1093/gbe/evt064

Sharma, S. K., Yamamoto, M., & Mukai, Y. (2018). Delineation of methylation and histone modification : The epigenetic regulatory marks show slightly altered distribution with the elevation in ploidy level in the orchid Dendrobium nobile. *The Nucleus*, *61*(3), 183-193. https://doi.org/10.1007/s13237-018-0231-1

Shewry, P. R., & Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food and Energy Security*, *4*(3), 178-202. https://doi.org/10.1002/fes3.64

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO : Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B | Genome Biology | Full Text. (s. d.). Consulté 14 juillet 2020, à l'adresse https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0754-6

Soltis, D. E., Visger, C. J., & Soltis, P. S. (2014). The polyploidy revolution then…and now : Stebbins revisited. *American Journal of Botany*, *101*(7), 1057-1078. https://doi.org/10.3732/ajb.1400178

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, *60*, 561-588. https://doi.org/10.1146/annurev.arplant.043008.092039

Song, Q., & Chen, Z. J. (2015). Epigenetic and developmental regulation in plant polyploids. *Current opinion in plant biology*, *24*, 101-109. https://doi.org/10.1016/j.pbi.2015.02.007

Song, Q., Zhang, T., Stelly, D. M., & Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biology*, *18*(1), 99. https://doi.org/10.1186/s13059-017-1229-8

Springer, N. M., Lisch, D., & Li, Q. (2016). Creating Order from Chaos : Epigenome Dynamics in Plants with Complex Genomes. *The Plant Cell*, *28*(2), 314-325. https://doi.org/10.1105/tpc.15.00911

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S. V., Obermeier, C., Parkin, I. A. P., Chèvre, A.-M., & Snowdon, R. J. (2017). Mapping of homoeologous

chromosome exchanges influencing quantitative trait variation in Brassica napus. *Plant Biotechnology Journal*, *15*(11), 1478-1489. https://doi.org/10.1111/pbi.12732

Strålfors, A., & Ekwall, K. (2011). Heterochromatin and Euchromatin—Organization, Boundaries, and Gene Regulation. In *Reviews in Cell Biology and Molecular Medicine*. American Cancer Society. https://doi.org/10.1002/3527600906.mcb.200400018.pub2

Szadkowski, E., Eber, F., Huteau, V., Lodé, M., Huneau, C., Belcram, H., Coriton, O., Manzanares-Dauleux, M. J., Delourme, R., King, G. J., Chalhoub, B., Jenczewski, E., & Chèvre, A.-M. (2010). The first meiosis of resynthesized Brassica napus, a genome blender. *New Phytologist*, *186*(1), 102-112. https://doi.org/10.1111/j.1469-8137.2010.03182.x

Tayalé, A., & Parisod, C. (2013). Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenetic and Genome Research*, *140*(2-4), 79-96. https://doi.org/10.1159/000351318

Udall, J. A., Quijada, P. A., & Osborn, T. C. (2005). Detection of Chromosomal Rearrangements Derived From Homeologous Recombination in Four Mapping Populations of Brassica napus L. *Genetics*, *169*(2), 967-979. https://doi.org/10.1534/genetics.104.033209

Ueda, M., & Seki, M. (2020). Histone Modifications Form Epigenetic Regulatory Networks to Regulate Abiotic Stress Response1[OPEN]. *Plant Physiology*, *182*(1), 15-26. https://doi.org/10.1104/pp.19.00988

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017a). The evolutionary significance of polyploidy. *Nature Reviews. Genetics*, *18*(7), 411-424. https://doi.org/10.1038/nrg.2017.26

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017b). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411-424. https://doi.org/10.1038/nrg.2017.26

Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research*, *24*(8), 1334-1347. https://doi.org/10.1101/gr.168997.113

Vergara, Z., & Gutierrez, C. (2017). Emerging roles of chromatin in the maintenance of genome organization and function in plants. *Genome Biology*, *18*(1), 96. https://doi.org/10.1186/s13059-017-1236-9

Vicient, C. M., & Casacuberta, J. M. (2017a). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, *120*(2), 195-207. https://doi.org/10.1093/aob/mcx078

Vicient, C. M., & Casacuberta, J. M. (2017b). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, *120*(2), 195-207. https://doi.org/10.1093/aob/mcx078

Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D., & Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(15), 4672-4677. https://doi.org/10.1073/pnas.1504783112

Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R. W., Comai, L., & Chen, Z. J. (2006). Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. *Genetics*, *172*(1), 507-517. https://doi.org/10.1534/genetics.105.047894

Weiss-Schneeweiss, H., Emadzade, K., Jang, T.-S., & Schneeweiss, G. M. (2013). Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. *Cytogenetic and genome research*, *140*(0). https://doi.org/10.1159/000351727

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *American Journal of Botany*, *102*(11), 1753-1756. https://doi.org/10.3732/ajb.1500320

Wertheim, B., Beukeboom, L. W., & Zande, L. van de. (2013). Polyploidy in Animals : Effects of Gene Expression on Sex Determination, Evolution and Ecology. *Cytogenetic and Genome Research*, *140*(2-4), 256-269. https://doi.org/10.1159/000351998

West, P. T., Li, Q., Ji, L., Eichten, S. R., Song, J., Vaughn, M. W., Schmitz, R. J., & Springer, N. M. (2014). Genomic Distribution of H3K9me2 and DNA Methylation in a Maize Genome. *PLOS ONE*, *9*(8), e105267. https://doi.org/10.1371/journal.pone.0105267

Wiles, E. T., & Selker, E. U. (2017). H3K27 methylation : A promiscuous repressive chromatin mark. *Current opinion in genetics & development*, *43*, 31-37. https://doi.org/10.1016/j.gde.2016.11.001

Winkler, H. (1917). Über die experimentelle Erzeugung yon Pflanzen mit abweichenden Chromosomenzahlen. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, *17*(3), 270-272. https://doi.org/10.1007/BF01740617

Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., & Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences*, *111*(14), 5283-5288. https://doi.org/10.1073/pnas.1402475111

Wu, J., Lin, L., Xu, M., Chen, P., Liu, D., Sun, Q., Ran, L., & Wang, Y. (2018). Homoeolog expression bias and expression level dominance in resynthesized allopolyploid Brassica napus. *BMC Genomics*, *19*(1), 586. https://doi.org/10.1186/s12864-018-4966-5

X, Q., H, A., Te, H., C, D., Pd, B., Jc, P., & Ms, B. (2019). *Genes derived from ancient polyploidy have higher genetic diversity and are associated with domestication in Brassica rapa*. https://doi.org/10.1101/842351

Xiong, Z., Gaeta, R. T., & Pires, J. C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. *Proceedings of the National Academy of Sciences*, *108*(19), 7908-7913. https://doi.org/10.1073/pnas.1014138108

Yoo, M.-J., Szadkowski, E., & Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, *110*(2), 171-180. https://doi.org/10.1038/hdy.2012.94

Zhang, D., Pan, Q., Cui, C., Tan, C., Ge, X., Shao, Y., & Li, Z. (2015). Genome-specific differential gene expressions in resynthesized Brassica allotetraploids from pair-wise crosses of three cultivated diploids revealed by RNA-seq. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00957

Zhang, D., Pan, Q., Tan, C., Zhu, B., Ge, X., Shao, Y., & Li, Z. (2016). Genome-Wide Gene Expressions Respond Differently to A-subgenome Origins in Brassica napus Synthetic Hybrids and Natural Allotetraploid. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01508

Zhang, K., Wang, X., & Cheng, F. (2019). Plant Polyploidy : Origin, Evolution, and Its Influence on Crop Domestication. *Horticultural Plant Journal*, *5*(6), 231-239. https://doi.org/10.1016/j.hpj.2019.11.003

Zhang, L.-Q., Liu, D.-C., Zheng, Y.-L., Yan, Z.-H., Dai, S.-F., Li, Y.-F., Jiang, Q., Ye, Y.-Q., & Yen, Y. (2010). Frequent occurrence of unreduced gametes in Triticum turgidum–Aegilops tauschii hybrids. *Euphytica*, *172*(2), 285-294. https://doi.org/10.1007/s10681-009-0081-7

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C. A., Scheffler, B. E., Stelly, D. M., Hulse-Kemp, A. M., Wan, Q., Liu, B., Liu, C., Wang, S., Pan, M., Wang, Y., Wang, D., Ye, W., … Chen, Z. J. (2015). Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnology*, *33*(5), 531-537. https://doi.org/10.1038/nbt.3207

Zhang, W., Fan, X., Gao, Y., Liu, L., Sun, L., Su, Q., Han, J., Zhang, N., Cui, F., Ji, J., Tong, Y., & Li, J. (2017). Chromatin modification contributes to the expression divergence of three TaGS2 homoeologs in hexaploid wheat. *Scientific Reports*, *7*. https://doi.org/10.1038/srep44677

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., Magdelenat, G., Couloux,

A., Samain, S., Gill, B. S., Rasmussen, J. B., Barbe, V., Faris, J. D., & Chalhoub, B. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1110552108

Zhao, M., Zhang, B., Lisch, D., & Ma, J. (2017). Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *The Plant Cell*, *29*(12), 2974-2994. https://doi.org/10.1105/tpc.17.00595

Zheng, D., Ye, W., Song, Q., Han, F., Zhang, T., & Chen, Z. J. (2016). Histone Modifications Define Expression Bias of Homoeologous Genomes in Allotetraploid Cotton. *Plant Physiology*, *172*(3), 1760-1771. https://doi.org/10.1104/pp.16.01210

Zhou, L., & Gui, J. (2017). Natural and artificial polyploids in aquaculture. *Aquaculture and Fisheries*, *2*(3), 103-111. https://doi.org/10.1016/j.aaf.2017.04.003

Zhu, W., Hu, B., Becker, C., Doğan, E. S., Berendzen, K. W., Weigel, D., & Liu, C. (2017). Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biology*, *18*. https://doi.org/10.1186/s13059-017-1281-4

Chapitre IV : Résultats méthodologiques

Amin, V., Harris, R. A., Onuchic, V., Jackson, A. R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C., & Milosavljevic, A. (2015). Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nature Communications*, *6*(1), 6370. https://doi.org/10.1038/ncomms7370

Baker, K., Dhillon, T., Colas, I., Cook, N., Milne, I., Milne, L., Bayer, M., & Flavell, A. J. (2015). Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance. *The Plant Journal: For Cell and Molecular Biology*, *84*(1), 111-124. https://doi.org/10.1111/tpj.12963

Bancel, E., Bonnot, T., Davanture, M., Branlard, G., Zivy, M., & Martre, P. (2015). Proteomic Approach to Identify Nuclear Proteins in Wheat Grain. *Journal of Proteome Research*, *14*(10), 4432-4439. https://doi.org/10.1021/acs.jproteome.5b00446

Chen, T., & Dent, S. Y. R. (2014). Chromatin modifiers and remodellers : Regulators of cellular differentiation. *Nature Reviews. Genetics*, *15*(2), 93-106. https://doi.org/10.1038/nrg3607

Debode, F., Marien, A., Gérard, A., Francis, F., Fumière, O., & Berben, G. (2017). Development of real-time PCR tests for the detection of Tenebrio molitor in food and feed. *Food Additives & Contaminants. Part A, Chemistry, Analysis, Control, Exposure & Risk Assessment*, *34*(8), 1421-1426. https://doi.org/10.1080/19440049.2017.1320811

Gent, J. I., Dong, Y., Jiang, J., & Dawe, R. K. (2012). Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications. *Nucleic Acids Research*, *40*(4), 1550-1560. https://doi.org/10.1093/nar/gkr862

How many biological replicates needed for ChIP seq experiments? (s. d.). Consulté 14 juillet 2020, à l'adresse https://www.biostars.org/p/274435/

Li, X., Tian, J., Nguyen, T., & Shen, W. (2008). Paper-Based Microfluidic Devices by Plasma Treatment. *Analytical Chemistry*, *80*(23), 9131-9134. https://doi.org/10.1021/ac801729t

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W., & Su, Z. (2018). PCSD : A plant chromatin state database. *Nucleic Acids Research*, *46*(D1), D1157-D1167. https://doi.org/10.1093/nar/gkx919

Makarevitch, I., Eichten, S. R., Briskine, R., Waters, A. J., Danilevskaya, O. N., Meeley, R. B., Myers, C. L., Vaughn, M. W., & Springer, N. M. (2013). Genomic Distribution of Maize Facultative Heterochromatin Marked by Trimethylation of H3K27[W]. *The Plant Cell*, *25*(3), 780-793. https://doi.org/10.1105/tpc.112.106427

Raha, D., Hong, M., & Snyder, M. (2010). ChIP-Seq : A Method for Global Identification of

Regulatory Elements in the Genome. *Current Protocols in Molecular Biology*, *91*(1), 21.19.1-21.19.14. https://doi.org/10.1002/0471142727.mb2119s91

Service My CoRe—My CoRe, partage et nomadisme. (s. d.). Service My CoRe. Consulté 14 juillet 2020, à l'adresse https://mycore.core-cloud.net/index.php/s/wMlm3VIIZdcnfwe

Shi, J., & Dawe, R. K. (2006). Partitioning of the maize epigenome by the number of methyl groups on histone H3 lysines 9 and 27. *Genetics*, *173*(3), 1571-1583. https://doi.org/10.1534/genetics.106.056853

Stark, R., & Hadfield, J. (2016). Characterization of DNA-Protein Interactions : Design and Analysis of ChIP-Seq Experiments. In A. M. Aransay & J. L. Lavín Trueba (Éds.), *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing* (p. 223-260). Springer International Publishing. https://doi.org/10.1007/978-3-319-31350-4_10

UCSC Genome Browser Home. (s. d.). Consulté 14 juillet 2020, à l'adresse http://systemsbiology.cau.edu.cn/

Widman, N., Feng, S., Jacobsen, S. E., & Pellegrini, M. (2014). Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. *Epigenetics*, *9*(2), 236-242. https://doi.org/10.4161/epi.26869

Chapitre V : Discussion Générale

Arthur, R. K., Ma, L., Slattery, M., Spokony, R. F., Ostapenko, A., Nègre, N., & White, K. P. (2014). Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. *Genome Research*, *24*(7), 1115-1124. https://doi.org/10.1101/gr.162008.113

Baker, K., Dhillon, T., Colas, I., Cook, N., Milne, I., Milne, L., Bayer, M., & Flavell, A. J. (2015). Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance. *The Plant Journal: For Cell and Molecular Biology*, *84*(1), 111-124. https://doi.org/10.1111/tpj.12963

Baranello, L., Kouzine, F., Sanford, S., & Levens, D. (2016). ChIP bias as a function of cross-linking time. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *24*(2), 175-181. https://doi.org/10.1007/s10577-015-9509-1

Baulcombe, D. C., & Dean, C. (2014). Epigenetic Regulation in Plant Responses to the Environment. *Cold Spring Harbor Perspectives in Biology*, *6*(9). https://doi.org/10.1101/cshperspect.a019471

Berke, L., Sanchez-Perez, G. F., & Snel, B. (2012). Contribution of the epigenetic mark H3K27me3 to functional divergence after whole genome duplication in Arabidopsis. *Genome Biology*, *13*(10), R94. https://doi.org/10.1186/gb-2012-13-10-r94

Berke, L., & Snel, B. (2014). The histone modification H3K27me3 is retained after gene duplication and correlates with conserved noncoding sequences in Arabidopsis. *Genome Biology and Evolution*, *6*(3), 572-579. https://doi.org/10.1093/gbe/evu040

Bottani, S., Zabet, N. R., Wendel, J. F., & Veitia, R. A. (2018). Gene Expression Dominance in Allopolyploids : Hypotheses and Models. *Trends in Plant Science*, *23*(5), 393-402. https://doi.org/10.1016/j.tplants.2018.01.002

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., & Tyler, J. K. (2016). The Overlooked Fact : Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Molecular and Cellular Biology*, *36*(5), 662-667. https://doi.org/10.1128/MCB.00970-14

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., Ruan, Y., Bickel, P. J., Myers, R. M., Wold, B. J., White, K. P., Lieb, J. D., & Liu, X. S. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, *9*(6), 609-614. https://doi.org/10.1038/nmeth.1985

Chica, C., Louis, A., Roest Crollius, H., Colot, V., & Roudier, F. (2017). Comparative

epigenomics in the Brassicaceae reveals two evolutionarily conserved modes of PRC2-mediated gene regulation. *Genome Biology*, *18*(1), 207. https://doi.org/10.1186/s13059-017-1333-9

Coate, J. E., & Doyle, J. J. (2010). Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species : An example from a plant allopolyploid. *Genome Biology and Evolution*, *2*, 534-546. https://doi.org/10.1093/gbe/evq038

Consortium (IWGSC), T. I. W. G. S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, *345*(6194). https://doi.org/10.1126/science.1251788

Consortium (IWGSC), T. I. W. G. S., Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403). https://doi.org/10.1126/science.aar7191

Cosseau, C., Wolkenhauer, O., Padalino, G., Geyer, K. K., Hoffmann, K. F., & Grunau, C. (2017). (Epi)genetic Inheritance in Schistosoma mansoni : A Systems Approach. *Trends in Parasitology*, *33*(4), 285-294. https://doi.org/10.1016/j.pt.2016.12.002

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., Wincker, P., Quesneville, H., Feuillet, C., & Choulet, F. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0546-4

Das, P. M., Ramachandran, K., vanWert, J., & Singal, R. (2004). Chromatin immunoprecipitation assay. *BioTechniques*, *37*(6), 961-969. https://doi.org/10.2144/04376RV01

De Storme, N., & Geelen, D. (2013). Sexual polyploidization in plants – cytological mechanisms and molecular regulation. *The New Phytologist*, *198*(3), 670-684. https://doi.org/10.1111/nph.12184

Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes. (s. d.). Consulté 17 août 2020, à l'adresse https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0110357

Dong, Z., Yu, J., Li, H., Huang, W., Xu, L., Zhao, Y., Zhang, T., Xu, W., Jiang, J., Su, Z., & Jin, W. (2018). Transcriptional and epigenetic adaptation of maize chromosomes in Oat-Maize addition lines. *Nucleic Acids Research*, *46*(10), 5012-5028. https://doi.org/10.1093/nar/gky209

Duboule, D. (2019). Le génome et ses embryons. *La lettre du Collège de France*, *44*, 11. https://doi.org/10.4000/lettre-cdf.4204

Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., Bowman, M. J., Childs, K. L., Washburn, J. D., Schmitz, R. J., Smith, G. D., Pires, J. C., & Puzey, J. R. (2017). Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *The Plant Cell*, *29*(9), 2150-2167. https://doi.org/10.1105/tpc.17.00010

Fernández, A. F., Toraño, E. G., Urdinguio, R. G., Lana, A. G., Fernández, I. A., & Fraga, M. F. (2014). The epigenetic basis of adaptation and responses to environmental change : Perspective on human reproduction. *Advances in Experimental Medicine and Biology*, *753*, 97-117. https://doi.org/10.1007/978-1-4939-0820-2_6

Fawcett, J. A., Maere, S., & Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5737-5742. https://doi.org/10.1073/pnas.0900906106

Feschotte, C. (2008). The contribution of transposable elements to the evolution of regulatory networks. *Nature reviews. Genetics*, *9*(5), 397-405. https://doi.org/10.1038/nrg2337

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication : Tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*, *60*, 433-453. https://doi.org/10.1146/annurev.arplant.043008.092122

Gardiner, L.-J., Quinton-Tulloch, M., Olohan, L., Price, J., Hall, N., & Hall, A. (2015). A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology*, *16*(1), 273. https://doi.org/10.1186/s13059-015-0838-3

Guo, H., Jiao, Y., Tan, X., Wang, X., Huang, X., Jin, H., & Paterson, A. H. (2019). Gene duplication and genetic innovation in cereal genomes. *Genome Research*, *29*(2), 261-269. https://doi.org/10.1101/gr.237511.118

Heard, E., & Martienssen, R. A. (2014). Transgenerational Epigenetic Inheritance : Myths and Mechanisms. *Cell*, *157*(1), 95-109. https://doi.org/10.1016/j.cell.2014.02.045

Hennig, L., & Derkacheva, M. (2009). Diversity of Polycomb group complexes in plants : Same rules, different players? *Trends in Genetics*, *25*(9), 414-423. https://doi.org/10.1016/j.tig.2009.07.002

Huan, Q., Mao, Z., Chong, K., & Zhang, J. (2018). Global analysis of H3K4me3/H3K27me3 in Brachypodium distachyon reveals VRN3 as critical epigenetic regulation point in vernalization and provides insights into epigenetic memory. *New Phytologist*, *219*(4), 1373-1387. https://doi.org/10.1111/nph.15288

Klironomos, F. D., Berg, J., & Collins, S. (2013). How epigenetic mutations can affect genetic evolution : Model and mechanism. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *35*(6), 571-578. https://doi.org/10.1002/bies.201200169

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813-1831. https://doi.org/10.1101/gr.136184.111

Li, L., Briskine, R., Schaefer, R., Schnable, P. S., Myers, C. L., Flagel, L. E., Springer, N. M., & Muehlbauer, G. J. (2016). Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC Genomics*, *17*(1), 875. https://doi.org/10.1186/s12864-016-3194-0

Li, W., Lin, Y.-C., Li, Q., Shi, R., Lin, C.-Y., Chen, H., Chuang, L., Qu, G.-Z., Sederoff, R. R., & Chiang, V. L. (2014). A robust chromatin immunoprecipitation protocol for studying transcription factor-DNA interactions and histone modifications in wood-forming tissue. *Nature Protocols*, *9*(9), 2180-2193. https://doi.org/10.1038/nprot.2014.146

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W., & Su, Z. (2018). PCSD : A plant chromatin state database. *Nucleic Acids Research*, *46*(D1), D1157-D1167. https://doi.org/10.1093/nar/gkx919

Lokhande, S. D., Ogawa, K., Tanaka, A., & Hara, T. (2003). Effect of temperature on ascorbate peroxidase activity and flowering of Arabidopsis thaliana ecotypes under different light conditions. *Journal of Plant Physiology*, *160*(1), 57-64. https://doi.org/10.1078/0176-1617-00990

Makarevitch, I., Eichten, S. R., Briskine, R., Waters, A. J., Danilevskaya, O. N., Meeley, R. B., Myers, C. L., Vaughn, M. W., & Springer, N. M. (2013). Genomic Distribution of Maize Facultative Heterochromatin Marked by Trimethylation of H3K27[W]. *The Plant Cell*, *25*(3), 780-793. https://doi.org/10.1105/tpc.112.106427

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., & Olsen, O.-A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science (New York, N.Y.)*, *345*(6194), 1250092.

https://doi.org/10.1126/science.1250092
Mf, P., & B, L. (2019, février). *Intergenerational and transgenerational epigenetic inheritance in animals*. Nature Cell Biology; Nat Cell Biol. https://doi.org/10.1038/s41556-018-0242-9
Middleton, C. P., Senerchia, N., Stein, N., Akhunov, E. D., Keller, B., Wicker, T., & Kilian, B. (2014). Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PloS one*, 9(3), e85761. https://doi.org/10.1371/journal.pone.0085761
Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of Gene Duplication in Plants1[OPEN]. *Plant Physiology*, *171*(4), 2294-2316. https://doi.org/10.1104/pp.16.00523
Parvathaneni, R. K., Bertolini, E., Shamimuzzaman, M., Vera, D. L., Lung, P.-Y., Rice, B. R., Zhang, J., Brown, P. J., Lipka, A. E., Bass, H. W., & Eveland, A. L. (2020). The regulatory landscape of early maize inflorescence development. *Genome Biology*, *21*(1), 165. https://doi.org/10.1186/s13059-020-02070-8
Payá-Milans, M., Poza-Viejo, L., Martín-Uriz, P. S., Lara-Astiaso, D., Wilkinson, M. D., & Crevillén, P. (2019). Genome-wide analysis of the H3K27me3 epigenome and transcriptome in Brassica rapa. *GigaScience*, *8*(12). https://doi.org/10.1093/gigascience/giz147
Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, *20*(1), 38. https://doi.org/10.1186/s13059-019-1650-2
Raff, R. A. 1996. *The Shape of Life: Genes, Development, and the Evolution of Animal Form.* University of Chicago Press, Chicago
Ricardi, M. M., González, R. M., & Iusem, N. D. (2010). Protocol : Fine-tuning of a Chromatin Immunoprecipitation (ChIP) protocol in tomato. *Plant Methods*, *6*(1), 11. https://doi.org/10.1186/1746-4811-6-11
Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., … Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal*, *30*(10), 1928-1938. https://doi.org/10.1038/emboj.2011.103
Savic, D., Gertz, J., Jain, P., Cooper, G. M., & Myers, R. M. (2013). Mapping genome-wide transcription factor binding sites in frozen tissues. *Epigenetics & Chromatin*, *6*(1), 30. https://doi.org/10.1186/1756-8935-6-30
Schmid, M. W., Heichinger, C., Coman Schmid, D., Guthörl, D., Gagliardini, V., Bruggmann, R., Aluri, S., Aquino, C., Schmid, B., Turnbull, L. A., & Grossniklaus, U. (2018). Contribution of epigenetic variation to adaptation in Arabidopsis. *Nature Communications*, *9*(1), 4446. https://doi.org/10.1038/s41467-018-06932-5
Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., & Odom, D. T. (2009). ChIP-seq : Using high-throughput sequencing to discover protein–DNA interactions. *Methods*, *48*(3), 240-248. https://doi.org/10.1016/j.ymeth.2009.03.001
Skene, P. J., & Henikoff, S. (s. d.). A simple method for generating high-resolution maps of genome-wide protein binding. *eLife*, *4*. https://doi.org/10.7554/eLife.09225
Song, J., Henry, H., & Tian, L. (2020). Drought-inducible changes in the histone modification H3K9ac are associated with drought-responsive gene expression in Brachypodium distachyon. *Plant Biology (Stuttgart, Germany)*, *22*(3), 433-440. https://doi.org/10.1111/plb.13057
Song, Q., Zhang, T., Stelly, D. M., & Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biology*, *18*(1), 99. https://doi.org/10.1186/s13059-017-1229-8

The regulatory landscape of early maize inflorescence development | bioRxiv. (s. d.). Consulté 27 juillet 2020, à l'adresse https://www.biorxiv.org/content/10.1101/870378v1.full

Thiebaut, F., Hemerly, A. S., & Ferreira, P. C. G. (2019). A Role for Epigenetic Regulation in the Adaptation and Stress Responses of Non-model Plants. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00246

Vimont, N., Quah, F. X., Schöepfer, D. G., Roudier, F., Dirlewanger, E., Wigge, P. A., Wenden, B., & Cortijo, S. (2019). ChIP-seq and RNA-seq for complex and low-abundance tree buds reveal chromatin and expression co-dynamics during sweet cherry bud dormancy. *Tree Genetics & Genomes*, *16*(1), 9. https://doi.org/10.1007/s11295-019-1395-9

Visger, C. J., Wong, G. K.-S., Zhang, Y., Soltis, P. S., & Soltis, D. E. (2017). Divergent gene expression levels between diploid and autotetraploid Tolmiea (Saxifragaceae) relative to the total transcriptome, the cell, and biomass. *BioRxiv*, 169367. https://doi.org/10.1101/169367

Visger, C. J., Wong, G. K.-S., Zhang, Y., Soltis, P. S., & Soltis, D. E. (2019). Divergent gene expression levels between diploid and autotetraploid Tolmiea relative to the total transcriptome, the cell, and biomass. *American Journal of Botany*, *106*(2), 280-291. https://doi.org/10.1002/ajb2.1239

Wang, J., Marowsky, N. C., & Fan, C. (2014). Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0110357

Wang, J., Tao, F., Marowsky, N. C., & Fan, C. (2016). Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes1[OPEN]. *Plant Physiology*, *172*(1), 427-440. https://doi.org/10.1104/pp.16.01177

Wang, Xiangfeng, Elling, A. A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X. S., & Deng, X. W. (2009). Genome-Wide and Organ-Specific Landscapes of Epigenetic Modifications and Their Relationships to mRNA and Small RNA Transcriptomes in Maize. *The Plant Cell*, *21*(4), 1053-1069. https://doi.org/10.1105/tpc.109.065714

Wang, Xutong, Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L., Wendel, J. F., & Liu, B. (2017). Gene-body CG methylation and divergent expression of duplicate genes in rice. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-02860-4

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., De Oliveira, R., Mayer, K. F. X., Paux, E., Choulet, F., & International Wheat Genome Sequencing Consortium. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology*, *19*(1), 103. https://doi.org/10.1186/s13059-018-1479-0

Whittaker C, Dean C. The FLC Locus: A Platform for Discoveries in Epigenetics and Adaptation. Annu Rev Cell Dev Biol. 2017;33:555-575. doi:10.1146/annurev-cellbio-100616-060546

Xu, C., Nadon, B. D., Kim, K. D., & Jackson, S. A. (2018). Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant, Cell & Environment*, *41*(9), 2033-2044. https://doi.org/10.1111/pce.13127

You, Q., Yi, X., Zhang, K., Wang, C., Ma, X., Zhang, X., Xu, W., Li, F., & Su, Z. (2017). Genome-wide comparative analysis of H3K4me3 profiles between diploid and allotetraploid cotton to refine genome annotation. *Scientific Reports*, *7*(1), 9098. https://doi.org/10.1038/s41598-017-09680-6

Zhang, H., Zheng, R., Wang, Y., Zhang, Y., Hong, P., Fang, Y., Li, G., & Fang, Y. (2019). The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation. *Nucleic Acids Research*, *47*(15), 7857-7869. https://doi.org/10.1093/nar/gkz511

Zhang, W., Garcia, N., Feng, Y., Zhao, H., & Messing, J. (2015). Genome-wide histone acetylation correlates with active transcription in maize. *Genomics*, *106*(4), 214-220.

https://doi.org/10.1016/j.ygeno.2015.05.005
Zhao, Hainan, Zhang, W., Chen, L., Wang, L., Marand, A. P., Wu, Y., & Jiang, J. (2018). Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. *Plant Physiology*, *176*(4), 2789-2803. https://doi.org/10.1104/pp.17.01467
Zhao, Huimin, Li, H., Jia, Y., Wen, X., Guo, H., Xu, H., & Wang, Y. (2020). Building a Robust Chromatin Immunoprecipitation Method with Substantially Improved Efficiency. *Plant Physiology*, *183*(3), 1026-1034. https://doi.org/10.1104/pp.20.00392
Zheng, Daoshan, Trynda, J., Sun, Z., & Li, Z. (2019a). NUCLIZE for quantifying epigenome : Generating histone modification data at single-nucleosome resolution using genuine nucleosome positions. *BMC Genomics*, *20*(1), 541. https://doi.org/10.1186/s12864-019-5932-6
Zheng, Daoshan, Trynda, J., Sun, Z., & Li, Z. (2019b). NUCLIZE for quantifying epigenome : Generating histone modification data at single-nucleosome resolution using genuine nucleosome positions. *BMC Genomics*, *20*(1), 541. https://doi.org/10.1186/s12864-019-5932-6
Zheng, Dewei, Ye, W., Song, Q., Han, F., Zhang, T., & Chen, Z. J. (2016). Histone Modifications Define Expression Bias of Homoeologous Genomes in Allotetraploid Cotton1[OPEN]. *Plant Physiology*, *172*(3), 1760-1771. https://doi.org/10.1104/pp.16.01210
Zhu, W., Hu, B., Becker, C., Doğan, E. S., Berendzen, K. W., Weigel, D., & Liu, C. (2017).
Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biology*, *18*. https://doi.org/10.1186/s13059-017-1281-4

ANNEXES

Annexe 1. Solutions pour le protocole de ChIP-seq

Extraction Buffer 1 Pour 100ml

0,4M Sucrose 20ml de 2M
10mM Tris-HCl pH8 1ml de 1M
10mM MgCl2 1ml de 1M
5mM B-mercapto 35µl de 14,4M
Proteases Inhibiteurs 200µl
(dissoudre 1 tablet dans 1 ml)
H20 77,8ml

Extraction Buffer 2 Pour 100ml

0,25M Sucrose 12,5ml de 2M
10mM Tris-HCl pH8 1ml de 1M
10mM MgCl2 1ml de 1M
1% Triton X-100 5ml de 20%
5mM B-mercapto 35µl de 14,4M
Proteases Inhibiteurs 200µl
(dissoudre 1 tablet dans 1 ml)
H20 80,3ml

Extraction Buffer 3 Pour 100ml

1,7M Sucrose 85ml de 2M
10mM Tris-HCl pH8 1ml de 1M
2mM MgCl2 200µl de 1M
0,15% Triton X-100 750µl de 20%
5mM B-mercapto 35µl de 14,4M
Proteases Inhibiteurs 200ul
(dissoudre 1 tablet dans 1 ml)
H20 12,8ml

LiCl Wash Buffer Pour 50ml

0,25M LiCl 3,125ml de 4M
1% NP40 (Igepal CA-630) 5ml de 10%
1mM EDTA 100µl de 0,5M
10mM Tris-HCl pH8 500µl de 1M
H20 41,25ml

TE Buffer Pour 50ml

1mM EDTA 100µl de 0,5M
10mM Tris-HCl pH8 500µl de 1M
H20 49,4ml

Elution Buffer Pour 20ml

1% SDS 1ml de 20%
0,1M NaHCO3 0,168g
H20 19ml

Nuclei Lysis Buffer Pour 5ml

50mM Tris-HCl pH8 250µl de 1M
10mM EDTA 100µl de 0,5M
1% SDS 250µl de 20%
Proteases Inhibiteurs 100µl
H20 4,3ml

ChIP dilution Buffer Pour 100ml

1,1 % Triton X-100 5,5ml de 20%
1,2mM EDTA 240µl de 0,5M
16,7mM Tris-HCl pH8 1,67ml de 1M
167mM NaCl 3,34ml de 5M
Proteases Inhibiteurs 200µl
H20 89,05ml

Low Salt Wash Buffer Pour 50ml

150mM NaCl 1,5ml de 5M
0,1% SDS 250µl de 20%
1% Triton X-100 2,5ml de 20%
2mM EDTA 200µl de 0,5M
20mM Tris-HCl pH8 1ml de 1M
H20 44,3ml

High Salt Wash Buffer Pour 50ml

500mM NaCl 5ml de 5M
0,1% SDS 250µl de 20%
1% Triton X-100 2,5ml de 20%
2mM EDTA 200µl de 0,5M
20mM Tris-HCl pH8 1ml de 1M
H20 41,05ml

Annexe 2 Dosage Qbit des quantités d'ADN obtenues à l'issue d'une expérience de ChIP réalisée à l'IPS2.

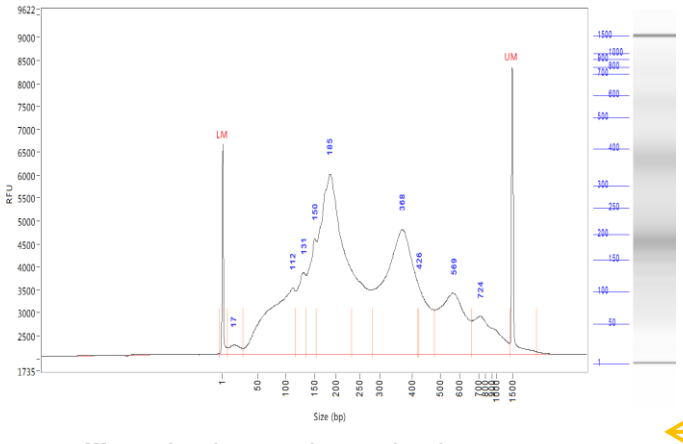
Les concentrations présentées correspondent à celles obtenues à l'issue de l'expérience de CHIP réalisée à l'IPS2 (colonne 2) puis celle obtenues après réalisation d'une deuxième immunoprécipitation sur la chromatine restante et non utilisée à Paris avec le même anticorps H3K27me3 utilisé (Colonne 3). La dernière colonne du tableau correspond à la quantité finale d'ADN présente dans les 20µL correspondant au pool des deux tubes obtenus pour chaque échantillon.

Stade de développement	Concentration IPS2 (ng/µl)	Concentration INRA (ng/µl)	Concentration totale ng (20µl)
Levée	0,12	0,3115	4,67
Levée	0,14	0,414	6,21
3 feuilles	0,19	0,322	4,83
3 feuilles	0,34	0,388	5,82
Tige 2 noeuds	0,27	0,748	8,98
Tige 2 noeuds	0,13	0,539	8,09
Epi méiose	0,95	0,474	7,11
Epi méiose	2,34	2,46	7,38
Epi floraison	0,76	0,687	8,93
Epi floraison	1,31	1,285	9,00
Grains 50°J	0,17	0,083	1,25
Grains 50°J	0,19	0,664	8,63
Grains 350°J	0,75	0,611	7,94
Grains 350°J	0,51	0,342	5,13
Grains 500°J	0,33	0,38	5,70
Grains 500°J	0,44	0,287	4,31
3 feuilles Renan	0,36	0,178	2,67
3 feuilles Renan	0,11	0,295	4,43
Input 3 feuilles CS	X	11,30	11,30
Input 3 feuilles Renan	10,50	26,60	26,60
Input Tg 2 noeud		x	x
Input Epi mé		15,20	15,20
Input G 50 °J		15,60	15,60

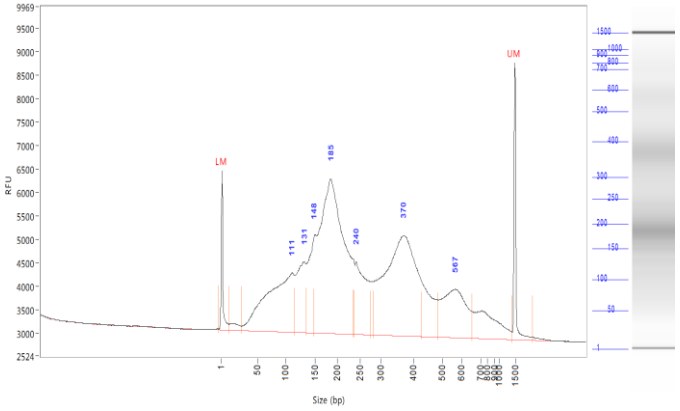
Annexe 3. Profils de Fragment Analyser des ADN obtenus à la suite de la fixation (15 minutes) et l'extraction de la chromatine pour différents temps de fragmentation pour le tissus feuille mature.

La flèche orange indique ~170pb (entre 150 et 200).

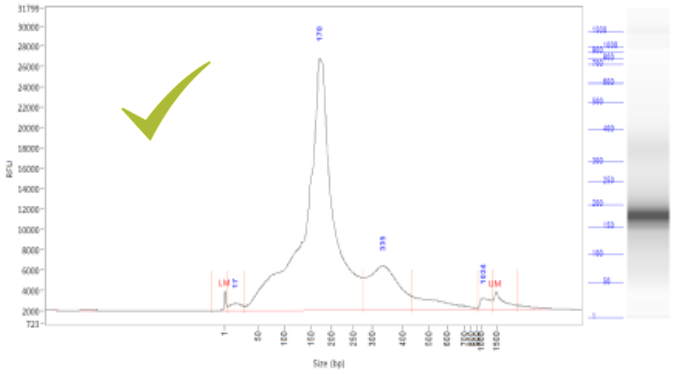
Feuilles Fixation15min Sonication 110 sec



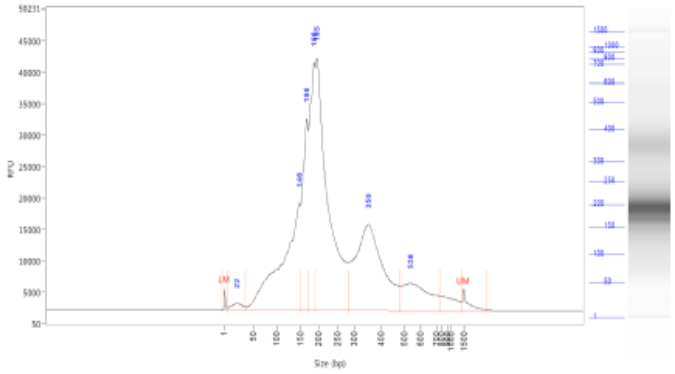
Feuilles Fixation15min Sonication 140 sec



Feuilles Fixation15min Sonication 220 sec

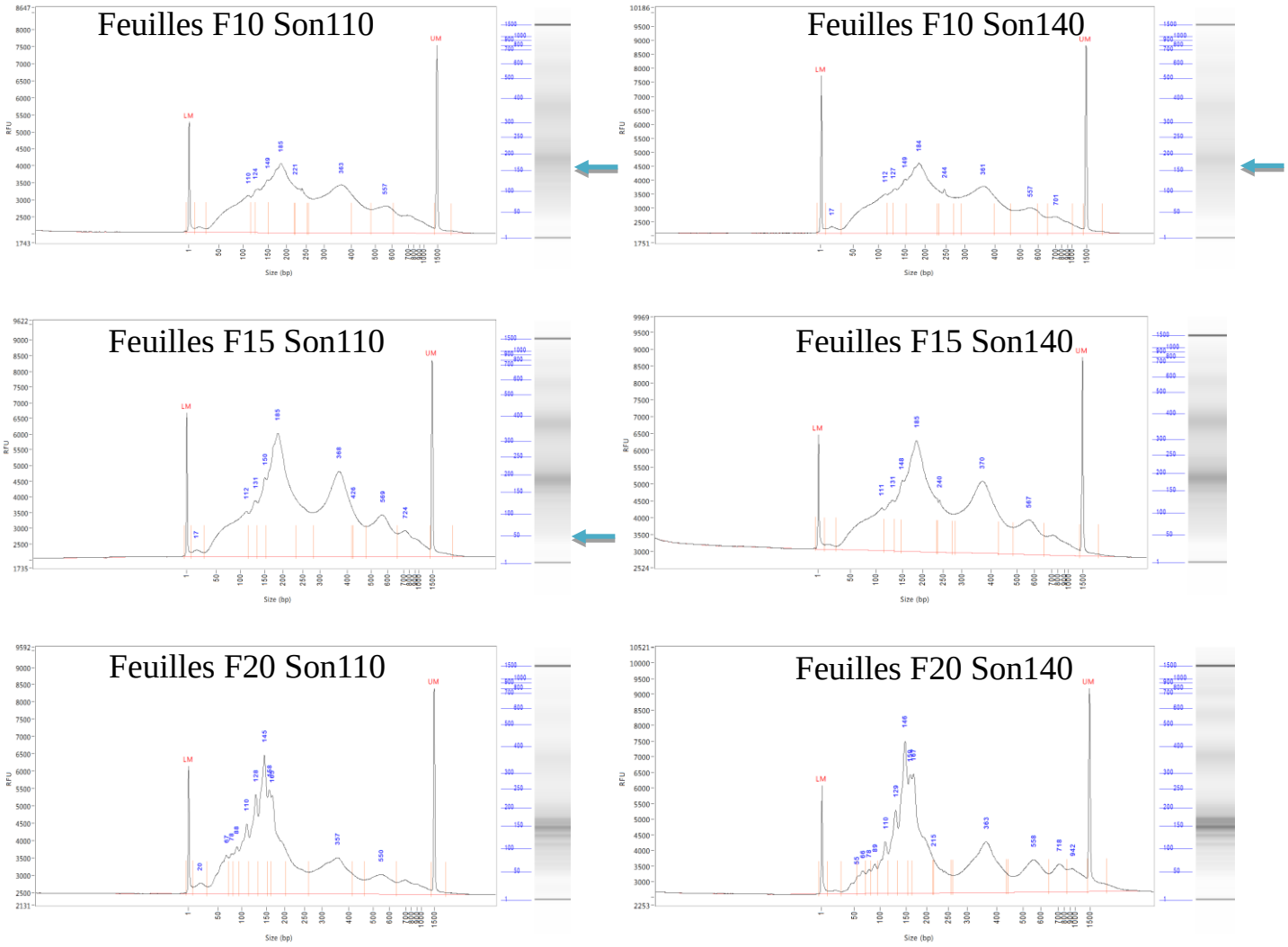


Feuilles Fixation15min Sonication 260 sec



Annexe 4. Profils de fragments d'ADN Fragment Analyzer pour différents temps de sonication sur feuilles fraiches matures pré-floraison

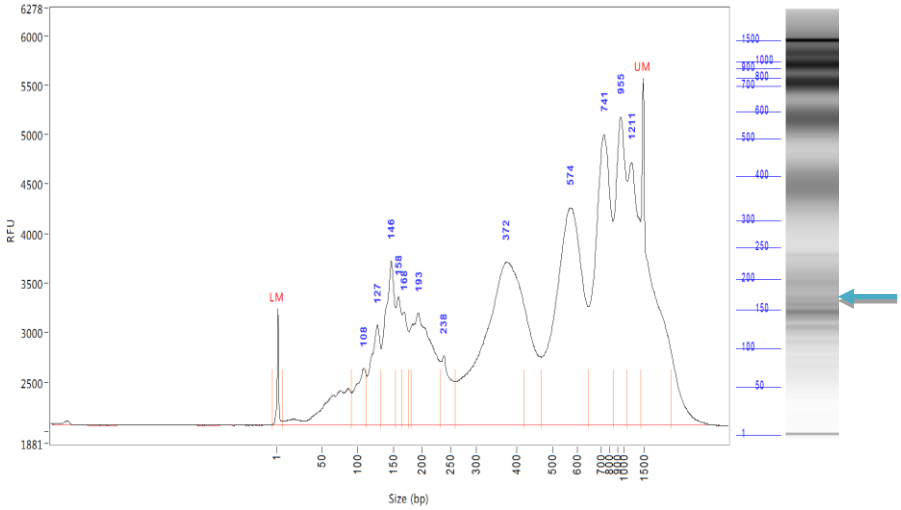
La flèche bleue représente 200pb sur le marqueur de taille.



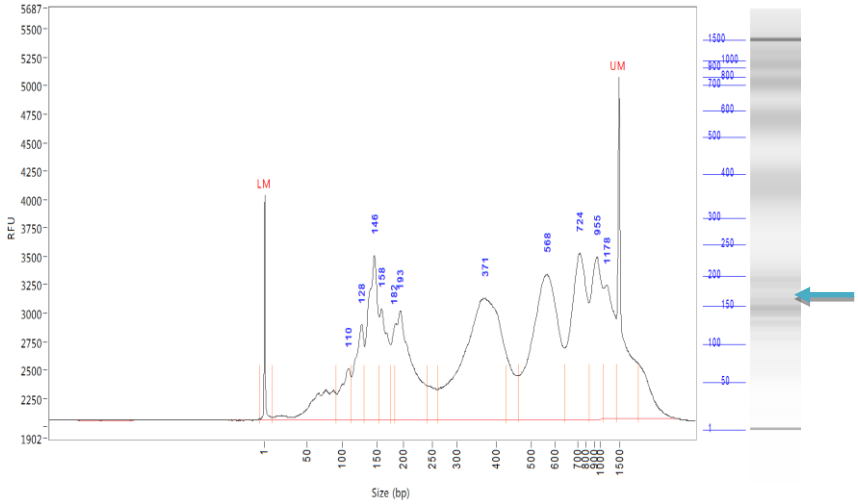
Annexe 5. Profils de fragments d'ADN Fragment Analyzer pour différents temps de sonication sur grains 500°J.

Flèche bleue = 200pb

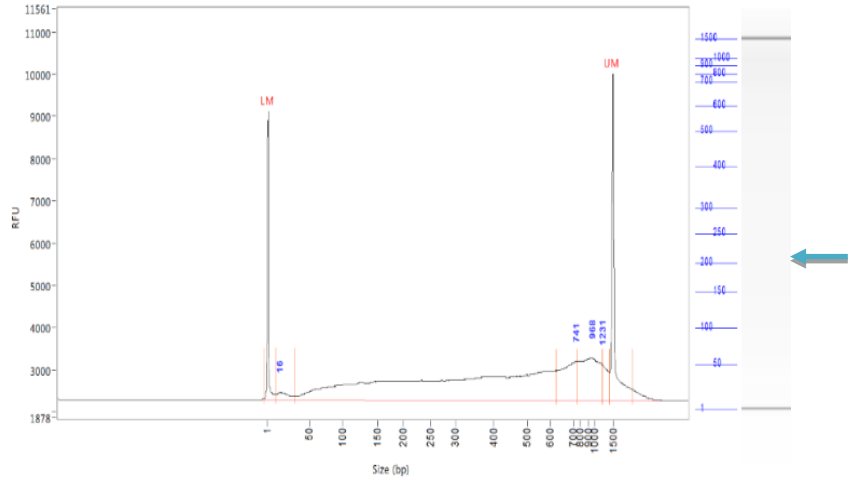
Grains Fixation :15 min Sonication:110 sec



Grains Fixation :15 min Sonication: 140 sec

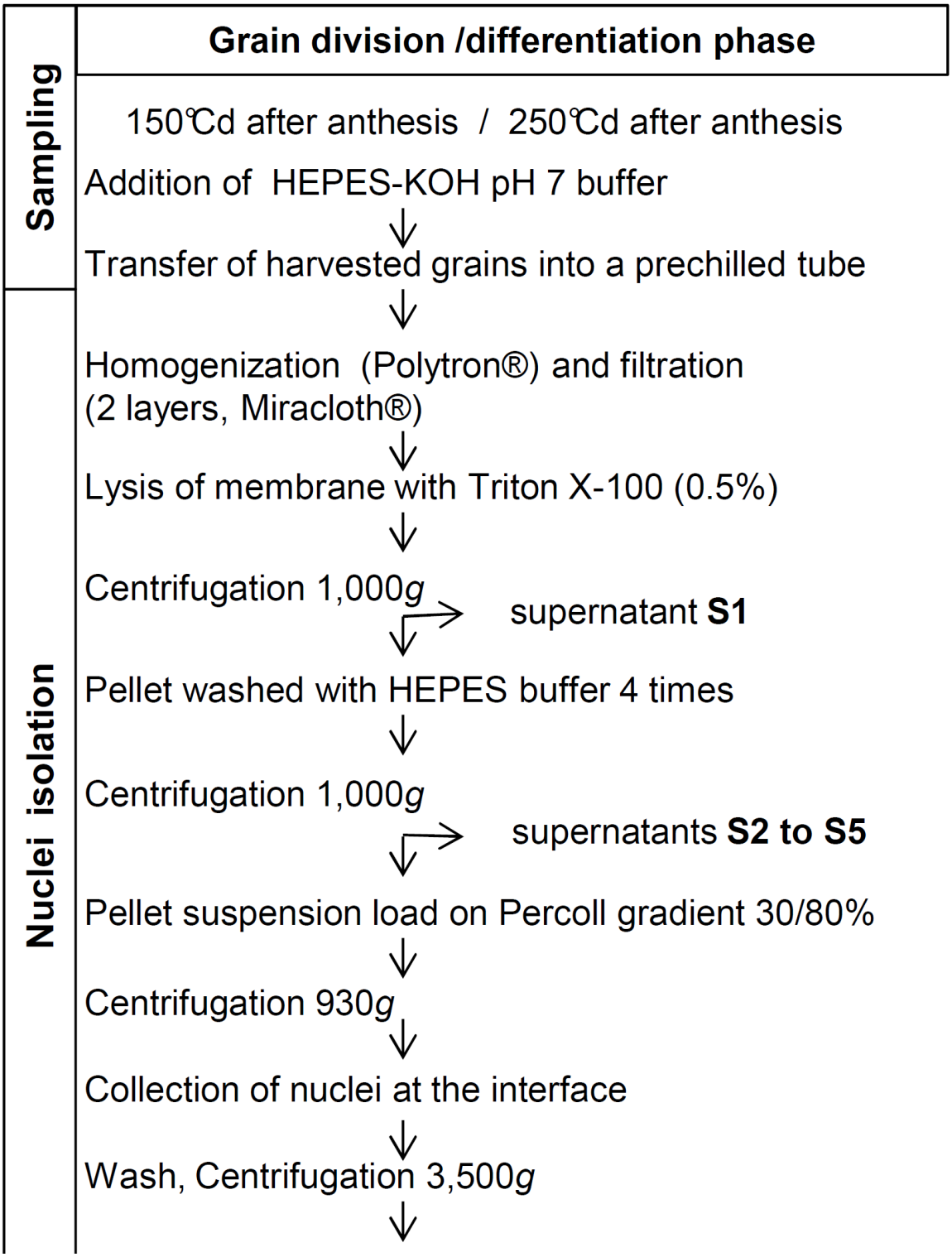


Grains Fixation :20 min Sonication: 110 sec

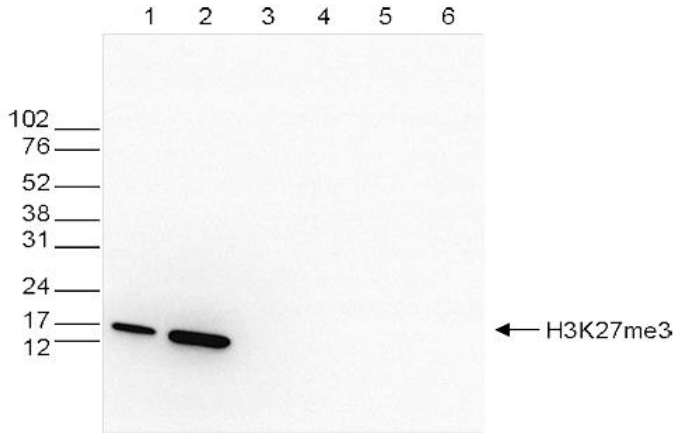


Annexe 6. Protocole pour isolation des noyaux des grains pour s'affranchir de l'amidon.

Bancel et al. 2015



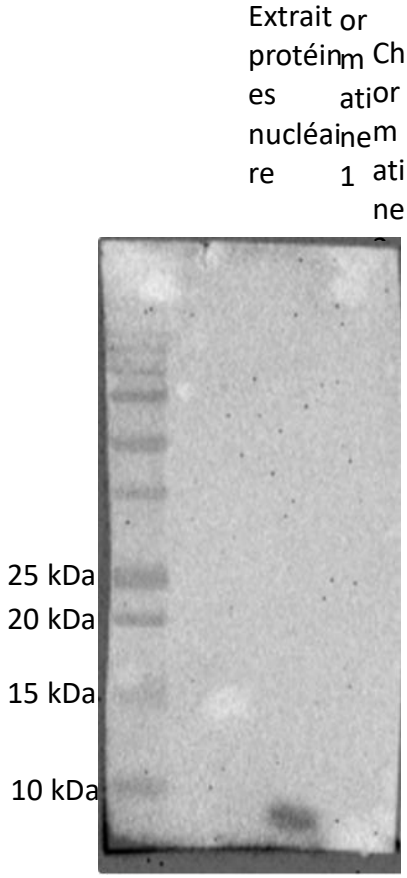
Annexe 7. Westernblot pour l'anticorps anti H3K27me3 de Diagenode



Westernblot obtenu par la société diagenode pour l'anticorps anti-H3K27me3.

1) 25µg sur un extrait de protéines protéines cellulaires de cellules HeLa
2) 15µg d'extraits d'histones
3) 1µg d'histones H2A ; 4) 1µg histones H2B ; 5) 1µg histones H3 ; 6) 1µg histones H4

Anticorps dilué au 1 :500 dans du TBS tween avec 5% de lait.



Westernblot obtenu au laboratoire pour l'anticorps de Diagenode anti H3K27me3

<u>Extraits protéiques nucléaires</u> (anthères F. Benyahya)

<u>Chromatine 1 non précipitée, stade 3 feuilles Chinese Spring</u>

80 µl extrait + 20 µl Laemli 5X

Ebullition 5 min

<u>Chromatine 2 précipitée (préparation Emanuelle Bancel):</u>

200µl extrait + 1ml acétone, 1h à -20°C

Centri 12 min 17000g

Culot lavé avec 1 ml acétone puis séché sous hotte

Culot + 100µl Laemli 1X, ébullition 5 min

<u>Dilution des anticorps</u>

Anticorps primaires dilués au 1/1000

Anticorps secondaire dilué au 1/100000 (rabbit)

Annexe 8. Protocole de Western Blot utilisé pour terster l'anicoprs utilisé pour les IP des études pilote.

1. Objet et domaine d'application

Un **Western blot**, ou immunoblot, est une méthode de biologie moléculaire qui permet la détection de protéines spécifiques d'un échantillon biologique et qui ont été transférées sur une membrane. Le **principe** consiste à appliquer sur la membrane un anticorps couplé à une HRP (Horseradish peroxidase) afin de révéler spécifiquement le(s) protéine(s) que l'on veut observer. La révélation se fait par chemiluminescence (kit ECL Clarity) par oxidation du luminol et émission de lumière.

2. Documents de référence

S'appuie sur un protocole de l'équipe biochimie qui utilise le kit ECL Clarity pour l'immunodétection et optimisé pour l'anticorps MUS 81.

3. Liste de diffusion et si nécessaire niveau de confidentialité

Document INRAE

4. Hygiène et sécurité

Port de gants, blouse. Manipulation des produits sous sorbonne. Pesée des produits CMR sur le poste de pesée sécurisé. Récupérer les déchets dans des poubelles spécifiques.

5. Matériels nécessaires

Mini-cuves d'électrophorèse Biorad (salle biochimie), générateur E2D, appareil de transfert, pipettes, balance de précision, appareil à eau ultra-pure.

6. Réactifs (chimiques et biologiques)

Produit	Fournisseur	Référence
Membrane de nitrocellulose 30 cm X 3,5 m	FISCHER Scientific	10551584
Tris-base	FISCHER Scientific	10376743
Chlorure de sodium	Elvetec	106401000
Acide chlorhydrique 37%	Sigma	258148-100mL
Tween 20	FISCHER Scientific	10485733
Lait 1/2 écrémé régilait en poudre	stock biochimie	Leclerc
Bovin serum Albumine heat schock fraction	MC2	A7906-100 g
Marqueur de taille Dual Color	Biorad	161-0374
AC primaire anti-rat de MUS81	Eurogentec	2 peptides de blé tendre dessinés par F. Benyahya
AC secondaire anti-rat couplé HRP fait chez la chèvre	Thermofischer Scientific	31475
ECL Clarity	Biorad	170-5061

7. Contraintes de la méthode

Travailler sous sorbonne

Durée totale de la manip : 5 heures, prévoir 1 H d'incubation pendant la pause déjeuner.

8. Contenu du mode opératoire

Immunodétection avec l'anticorps MUS81

Repérer sur la membrane les bandes du marqueur de taille au crayon à papier (après avoir coloré au Rouge ponceau pour vérifier la qualité du transfert des protéines et le repérage des tailles de(s) bandes).

Coloration de la membrane au Rouge de ponceau (témoin de transfert des protéines sur la membrane).

Rouge ponceau 0,1 g

Acide acétique 5 mL

Eau qsp 100 mL

- La membrane qui doit être colorée est glissée délicatement dans une petite boîte transparente préalablement étiquetée. Placer la boîte sous agitation lente pendant 5 min.

- Rincer 3X 5 min à l'eau osmosée sous agitation.
- Scanner la membrane (scanner BIORAD) et conserver une photo au format jpeg. Cette membrane servira de référence pour le repérage de taille de la protéine recherchée sur la membrane grâce au marqueur de taille.

Placer ensuite cette membrane (ou une autre ayant migré en même temps) dans les bains suivants (petites boîtes transparentes **recouvertes de papier aluminium**) :

- Saturation de la membrane : incubation 1 nuit à 4°C sous agitation dans TBS-Régilait 2.5 % (p/v) : 30 mL
- Lavages rapides : 2X 5 min dans TBS-Tween 0.05% (v/v) : 30 mL
- Hybridation avec AC MUS81 (1/500ème) : incubation 1H à T° ambiante sous agitation (40 µL dans 20 mL) dans TBS-Tween 0.05 %- Régilait 2.5 %.
- Vérifier que la membrane est bien recouverte par le produit et mobile.
- Lavages dans TBS-Tween 0.1 % (v/v) 3X 10 min (30 mL par membrane)
- Hybridation avec AC IIaire anti-rat HRP : incubation 1H à T° ambiante avec agitation dans 50 mL d'une dilution au 1/50.000ème (1 µL dans 50 mL de TBS—Tween 0.05%-BSA 5%).
- Lavages dans 30 mL de TBS-tween 0.05% (v/v) : 3X 10 min
- Lavages dans TBS seul 2X 15 min

Révélation avec le kit de détection ECL

Le kit se compose de 2 réactifs qu'il faut mélanger 1/1 (v/v) à raison de 0.125 mL/cm2 de membrane à traiter, soit 2 mL de chaque réactif pour une membrane de 5.4 x 8.2 cm = taille d'un mini gel.

Attention : sortir le kit 15 min avant utilisation.

- Incubation avec le mélange pendant 5 min en agitant la boîte ou dépôt direct sur la membrane placée sur une plaque de verre, sous cache aluminium.
- Eliminer l'excès de réactifs (récupération dans flacons à déchets) et enfermer la membrane dans du scellofrais.
- Placer la membrane dans la caméra et scanner.

Résumé de la manip :

Etape	Durée	Horaire (approximatif)
Saturation TBS-lait	1 nuit à 4°C	17 H
Lavages TBS-Tween-lait	2X 5 min	9h30-9h40
Hybridation AC Iaire	2H T° amb	9h-45-11h45
Lavages TBS-tween	3X 10 min	11h45-12h15
Hybridation AC Iiaire-HRP	1H	12h15-13h15
Lavages TBS-Tween-BSA	3X 10 min	13h15-13h45
Lavages TBS	2X 15 min	13h45-14h15

PREPARATION DES SOLUTIONS

Tampon TBS 1X pH= 7.6

Prévoir 2L pour 4 membranes

Tris-base 1.21 g (10 mM)

NaCl 8.766 g (150mM)

H2O up qsp 1000 mL

Ajuster à pH=7.6 avec du HCl fumant (environ 1.2 mL)

Conservation à 4°C

Tampon TBS-Tween 0.05%-Régilait 2.5%

Tampon TBS 1X pH 7.6 50 mL

Régilait 1.25 g

Conservation à 4°C

Tampon TBS-Tween 0.1%

Tampon TBS 1X pH 7.6 50 mL

Tween 20 50 µL

Tampon TBS-Tween 0.05%- BSA 5%

BSA 2.5 g

Tampon TBS 1X pH 7.6 50 mL

Tween 20 25 µL

Résumé

De nombreuses espèces de plantes sont polyploïdes, c'est-à-dire qu'elles possèdent plusieurs sous-génomes au sein du noyau de leurs cellules. La polyploïdie s'accompagne d'une redondance génétique qui offre un potentiel d'innovations évolutives important par un relâchement de la pression de sélection autorisant sous-fonctionnalisation, néo-fonctionnalisation, perte de gènes. Le blé tendre est une espèce polyploïde récente, apparue suite à deux hybridations interspécifiques (800 000 et 10 000 ans). Il possède un génome hexaploïde composé de trois sous-génomes : AABBDD et théoriquement, il possède trois copies homéologues de chaque gène (1A:1B:1D). Cependant, les analyses génomiques ont révélées que la moitié des séquences codantes présentaient un nombre de copie de type NA:NB:ND. Comment évolue cette redondance génétique après la polyploidisation chez le blé tendre? Peut-on observer des différences d'expression des copies de gènes témoignant d'une évolution fonctionnelle pour cette espèce formée très récement? Quels sont les mécanismes sous-jacents ?

L'objectif de cette thèse a été d'analyser les expressions relatives des copies de gènes homéologues pour des groupes présentant trois (1 :1 :1, triades), deux (0:1:1, 1:0:1 ou 1:1:0, dyades) ou quatre copies (2:1:1, 1:2:1 ou 1:1:2, tétrades). Nous avons également relié les résultats aux caractéristiques structurales (position génomique), évolutives (présence ou absence des copies chez les espèces ancêtres) et épigénétiques (marques histones) des gènes pour répondre aux questions de recherche. Nous avons utilisé les données de RNA-seq et de ChIP-seq mises à disposition lors de la publication dela séquence génomique de référence du blé tendre (IWGSC 2018).

Nous avons mis en évidence que les 51,1% de gènes en triades présentent en majorité (81%) une expression équilibrée sur l'ensemble des tissus et au cours du développement (expression élevée et constitutive). Ces gènes sont majoritairement associés la marque épigénétique d'activation de l'expression : H3K9ac. *A contrario,* les gènes en dyades (11,7% des gènes) et en tétrades (2,8% des gènes) présentent plus fréquemment des biais d'expression (36% et 75,4% respectivement). Ces gènes sont plus associés à la marque épigénétique liée à la répression ciblée et transitoire des séquences (H3K27me3). En revanche, aucune dominance d'expression n'a été décelée à l'échelle du génome entier. Ceci met en évidence de potentielles sous-fonctionnalisations des gènes, plus fréquentes pour des gènes différents des triades, présents dans les régions distales des chromosomes. Même si les biais d'expression correspondent à des différences déjà existantes chez les espèces ancêtres, nous avons cependant distingué des traits d'expression correspondant aux différentes étapes de l'histoire évolutive du blé : les copies du sous-génome D sont moins réprimées et moins associées à la marque H3K27me3 ; les biais d'expression entre les copies AABB sont plus prononcés. Ainsi, la coévolution des deux sous-génomes AABB pendant 800 000 ans est décelable alors que le sous-génome D semble encore s'exprimer de façon autonome.

Ces résultats suggèrent que ce génome comprend des gènes très contraints évolutivement qui constitueraient le « core » génome de l'espèce avec des fonctions de bases conservées (gènes en triades) et des gènes présentant des variations du nombre de copies, des régulations différentielles et des fonctions spécifiques témoignant de possibles innovations évolutives, appartenant probablement au génome dit « dispensable » (dyades et tétrades).

Mots clefs : polyploïdisation – *Triticum aestivum* – biais d'expression homéologues – épigénétique - évolution