Caractérisation de la reprogrammation de l'expression des gènes chez les blés allopyloïdes

Smahane Chalabi

UNIVERSITÉ EVRY VAL D'ESSONNE

Ecole Doctorale - Des Génomes Aux Organismes (GAO)

Laboratoire :
UNITE DE RECHERCHE EN GENOMIQUE VEGETALE
Equipe de recherche :
Organisation et évolution des génomes de plantes (OEGP)

Caractérisation de la reprogrammation de l'expression des gènes chez les blés allopolyploïdes

Thèse de Doctorat
Spécialité : Bioinformatique

Présentée et soutenue à Evry, le 13 Novembre 2014, par

Smahane CHALABI

COMPOSITION DU JURY

François ARTIGUENAVE	Directeur de Recherche, CNG Evry	Rapporteur
Philippe LASHERMES	Directeur de Recherche, INRA Montpellier	Rapporteur
Claudine DEVAUCHELLE	Maître de Conférences, Université d'Evry	Examinateur
Armel SALMON	Maître de Conférences, INRA Rennes	Examinateur
Julien CHIQUET	Maître de Conférences, CNRS/INRA Evry	Co-Encadrant
Boulos CHALHOUB	Directeur de Recherche, INRA Evry	Directeur de thèse

Remerciements

Je présente ma profonde gratitude à Boulos Chalhoub, mon directeur de thèse, de m'avoir donné ma chance et de m'avoir fait confiance pour mener ces travaux de recherche. Je le remercie pour toute la patience et les conseils qu'il m'a apportés tout au long de ma thèse. Je remercie Julien Chiquet, mon co-encadrant, pour m'avoir encadré en statistique et surtout en programmation. Je tiens à lui présenter ma reconnaissance pour toute la patience qu'il a déployée.

Je remercie sincèrement les membres du jury de ma thèse, qui malgré un emploi du temps très chargé ont accepté de lire mon travail et d'en juger la qualité. Je tiens à remercier François Artiguenave et Philippe Lashermes, d'avoir acceptés d'être rapporteurs, et Armel Salmon et Claudine Devauchelle d'avoir acceptés d'examiner mes travaux de thèse.

Je remercie les membres de mon comité de thèse : Jean-Marc Aury et Valérie Geoffroy, pour leur disponibilité, la réflexion et la discussion scientifique que nous avons pu avoir ensemble.

Je remercie Flavio Toma (ex-directeur de l'IUP GBI) pour m'avoir ouvert les portes de la bioinformatique, et tous les enseignants de l'IUP GBI qui m'ont donné cette envie de poursuivre dans ce domaine.

Je remercie tout mes collègues et amis de l'URGV pour leur bonne humeur au quotidien et leur encouragements du G1 au G2 en passant par le G1+1, plus particulièrement le groupe bioinfo (Véro, Jean-Philippe, Marie-Laure, Guillem, Zakia pour ton adorable soutien psychologique en fin de journée, Rim, Cécile, Etienne), nos informaticiens: Fifi (merci pour ton fort intérêt à toujours nous faciliter dans l'espace de stockage des données ou de calculs sur les serveurs) et Jean-Luc, les filles d'à coté de la plateforme transcriptomique (Stéphanie et Stéphanie, Ludivine, Alex, Sandrine..), le groupe de Claire… Je remercie aussi le service administratif de l'URGV (Mélanie, Arnaud, Sabine, Louisa) et de l'Université d'Evry (Saadia Diani, Florence Hamon, Véronique Fournie, Carole Troussier).

Je remercie, très chaleureursement, mon équipe pour les discussions scientifiques menées ensemble, pour son soutien au quotidien et les agréables moments partagés ensemble : Harry et les mignardises de Katia ou le chocolat, Vinh Ha et le chocolat Suisse, Nathalie et les p'tits biscuits, Isa et les balades en footing à la pause déjeuner ou les bonnes crêpes. Ce sont des très moments que j'ai partagé et qui m'ont beaucoup apporté moralement pour la bonne réalisation de ma thèse. Merci aux thésards qui m'ont précédée et les postdoc, qui ont travaillé dans l'équipe: Houda, JJ, Mathieu, Imen, Dominique, Edith, Cléa, Cyril. Un grand Merci à Houda, JJ, et Edith pour leurs précieux conseils et les discussions scientifiques très enrichissantes lors de mes travaux de thèses.

Un grand merci à Corinne DaSilva, qui m'a énormément appris sur l'alignement des lectures sur les serveurs du Genoscope. Je lui présente mes profonds remerciements pour sa disponibilité malgré une surcharge de travail, pour les discussions très fructueuses, pour ses conseils… Je remercie aussi l'équipe informatique du genoscope.

Je présente mes remerciements à toute l'équipe du Laboratoire de Statistique et Génome, pour leur aimabilité lorsque j'ai été amené à travailler avec Julien en statistique ou encore avec Claudine en bioinformatique. Je remercie énormément Carène de m'avoir encouragé à mener cette thèse avec Boulos Chalhoub, et pour la fructueuse collaboration dans le cadre du projet ANR où s'inscrivent mes travaux de thèse. Je remercie Claudine Devauchelle, pour toute sa patience et tout le temps qu'elle m'a offert lors de la programmation en python. Aussi, je n'oublie pas les stagiaires qui ont participé à ce projet ANR : Xi Liu et David Christiani.

Pour finir, je remercie profondément ma famille qui m'a plus que soutenue et encouragée tout au long de mes études, et encore plus pendant mon doctorat. Je remercie énormement mes soeurs Samira, Hasna et Naoual, ma belle-soeur Meriem qui n'ont jamais cessé de m'encourager pour ne pas lâcher prise. Un grand Merci à mon frère Faouzi, qui m'a beaucoup conseillé et motivé à entreprendre ma thèse et à donner du meilleur de moi-même. Merci à Nasser, qui m'a aussi toujours soutenu et encouragé dans mes choix, et pour sa patience durant ces quatre ans. Ma petite princesse Wassila est née au cours de la deuxième année de thèse, bien que ce ne fût

pas toujours évident entre les nuits blanches et le travail le lendemain, elle a été un véritable moteur dans cette réussite. Mes neveux Adam, Nadhir, Rayan et mes nièces Soukaïna et Omayma, je vous remercie pour tout le bonheur que vous m'offrez quand je vous vois!

Maman, comment te remercier suffisamment, tu as toujours été là pour moi, pour tout...je ne te remercierai jamais assez pour tout ce que tu m'as offert. Mon Père, à toi, je te dédie cette thèse, car dans les moments les plus difficiles à surmonter je pense à toi...

Table des matières

Liste des Figures et Tableau

Abréviations

ADN ou DNA: acide désoxy-ribonucléique

ADNc ou cDNA: ADN complémentaire

ADNr ou rDNA: ADN ribosomique

ARN ou RNA: acide ribonucléique

ARNdb ou dsRNA: ARN double-brin

ARNm ou mRNA: ARN messager

ARNr ou rRNA: ARN ribosomique

ATP: adénosine tri-phosphate

BAC : bacterial artificial chromosome

cDNA-AFLP: cDNA Amplified fragment length polymorphism

cv : cultivar

dNTP : désoxynucléotide tri-phosphate

EST : expressed sequence tag

ET : élément transposable

FDR : First Division Restitution (absence de première division de méiose)

HNRT: Homoeologous Non-Reciprocal Translocation

HE: Homoeologous Exchange

hpRNA: hairpin RNA

kb : kilobase

LTR : long terminal repeat

Ma / Mya : million d'année / million years ago

Mb : mégabase

miRNA : micro ARN

NGS : Next Generation Sequencing

pb : paire de bases

PCR : polymerase chain reaction

Ph1 : Pairing homeologous 1 (locus contrôlant l'appariement homéologue chez le blé)

PTGS : post-transcriptional gene silencing

RNAi : ARN interférence

RT : reverse-transcription

SDR : Second Division Restitution (absence de seconde division de méiose)

siRNA : small interfering RNA (petits ARN interférents)

ssp : sub-species en anglais (sous-espèce)

miRNA : micro ARN interférents

Préambule

La polyploïdie ou la duplication des génomes est une force majeure dans l'évolution et l'adaptation des espèces, notamment chez les plantes. Les raisons de ce succès évolutif ne sont pas encore bien élucidées et font l'objet de nombreuses recherches. Il a maintenant été montré que la polyploïdie induit une cascade de changements génétiques, fonctionnels et épigénétiques, générant de nouveaux phénotypes, différents de ceux des espèces parentales.

La polyploïdie est particulièrement importante chez les espèces de blé genres (*Triticum* et *Aegilops*) qui constituent, en plus de leur importance économique, un excellent modèle pour comprendre les mécanismes impliqués dans la stabilisation et le succès des espèces polyploïdes.
Mon travail de thèse a permis de caractériser la reprogrammation de l'expression des gènes et des homéologues qui les composent; en réponse à l'augmentation du niveau de ploïdie, comme déjà réalisé dans d'autres modèles polyploïdes, mais aussi, en réponse à la diminution du niveau de ploïdie, ce qui est particulier au modèle du blé étudié ici.

Ce manuscrit de thèse se divise en trois parties principales :

La première partie présente l'étude bibliographique de mon sujet de recherches. Cette partie comporte un seul chapitre : le **chapitre 1** qui fait « une synthèse » des recherches menées sur la polyploïdie et le blé afin de positionner mon sujet de recherche. Je présente, dans l'ordre, la polyploïdie et ses conséquences au niveau génétique, épigénétique et surtout l'expression des gènes, puis la présentation du modèle d'étude utilisé: le blé. Je poursuis ce chapitre par une présentation des outils d'étude du transcriptome, allant de la description des techniques utilisées depuis les premiers développements aux outils NGS (nouvelles techniques de séquençage), à l'analyse des données haut-débits générées. Enfin, je présente le séquençage du génome du blé, et je termine par la problématique de mon sujet de recherche.

La deuxième partie de mon manuscrit présente les résultats obtenus au cours de ma thèse, que j'ai rédigé, sous la forme d'articles qui sont présentés dans les chapitres 2, 3.et 4.
Le **chapitre 2** consiste en la caractérisation des changements de l'expression globale des gènes des blé allohexaploïdes génétiquement très stables, utilisant les puces Affymetrix, et montrant la prévalence de l'additivité. Ce travail a été publié dans la revue New Phytologist intitulé «Prevalence of gene expression additivity in genetically stable wheat allohexaploids».

Le **chapitre 3** est consacré à l'analyse des changements de l'expression des gènes en diminuant le niveau de polyploïdie (par l'extraction des tétraploïdes de blé à partir de blé hexaploïde) puis en la ré-augmentant, soit un retour à l'état hexaploïde. Une première tentative de disséquer l'expression globale des gènes en celle des sous-génomes qui les composent et ce avant l'ère des NGS est aussi présentée. Celle-ci consiste en l'élaboration de sondes oligonucléotidiques, parents spécifiques (PSF) et en analysant leur expression dans ce modèle. Ce chapitre est présenté sous forme d'article non encore soumis, intitulé « Unraveling gene expression changes when decreasing and re-increasing alloploidy in wheat ».

Le **chapitre 4** consiste en l'analyse, par le séquençage massif des ARNm, utilisant les outils NGS, des changements de l'expression des gènes dans ce modèle et son partitionnement en étudiant l'expression des homéologues qui les composent. Ce chapitre est également rédigé sous forme d'article, intitulé « Duplicate and Partitioning of homoeolog gene expression in the decreasing, re-increasing wheat polyploid model ».

La troisième et dernière partie de ce manuscrit présente le **chapitre 5**, c'est une discussion générale sur mes travaux de thèse, à laquelle je propose des perspectives.

Les **annexes** et **références bibliographiques** sont présentées à la fin du manuscrit.

Première Partie

Chapitre 1

Étude Bibliographique Générale

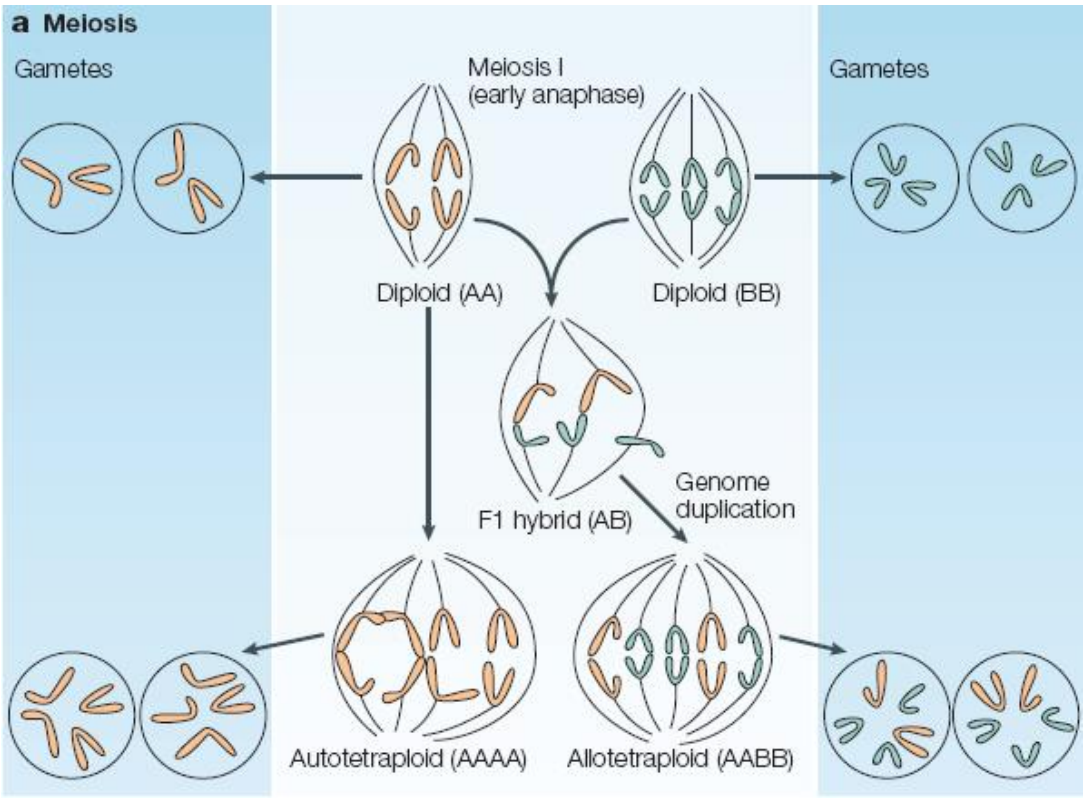


Figure 1: Formation des autopolyploïdes ou allopolyploïdes.

La figure illustre la composition et le comportement des chromosomes des diploïdes et de leur dérivés polyploïdes, d'après (Comai, 2005).

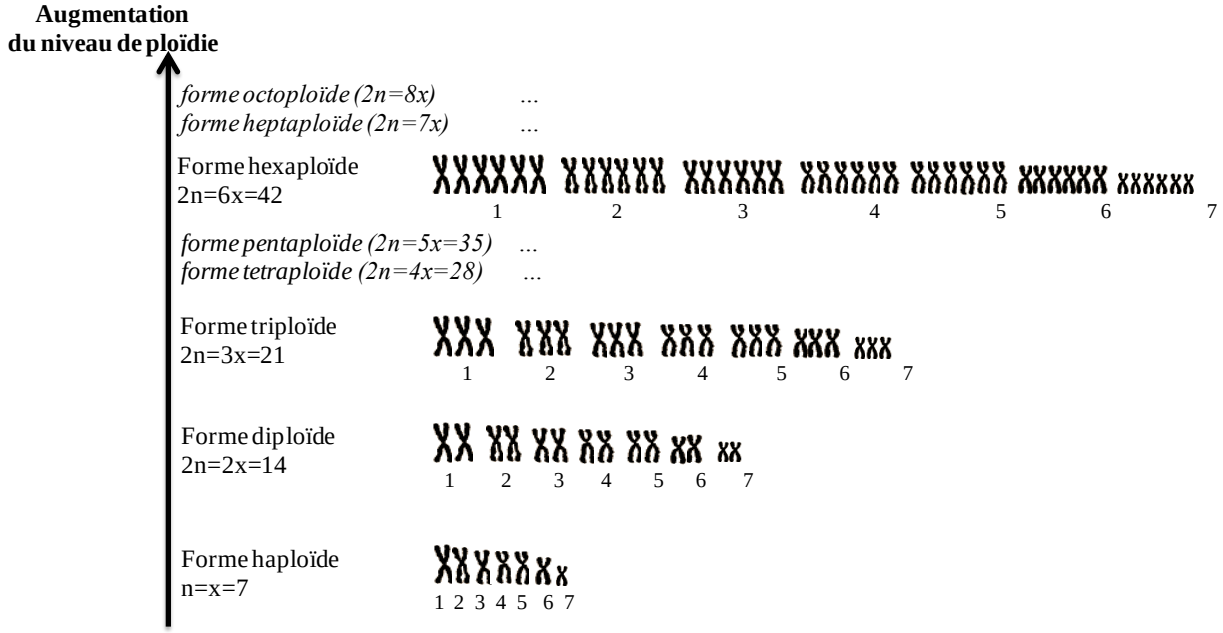


Figure 2: Schéma du nombre de chromosomes selon le niveau de ploïdie.

1.1. La polyploïdie

1.1.1. Définition

La polyploïdie est la duplication du génome entier, conduisant à l'assortiment de plusieurs jeux de chromosomes (supérieur à deux) dans un noyau.

La polyploïdie peut apparaître à l'échelle d'un organisme entier, d'un organe, d'un tissu ou même d'une simple cellule. On peut classer les individus polyploïdes selon l'origine de leur ploïdie : autopolyploïdie et allopolyploïdie (Fig. 1). Mais également suivant le niveau de ploïdie (triploïdie, tétraploïdie, pentaploïde…) (Fig. 2).

L'autopolyploïdie est la duplication du génome d'une même espèce : la duplication est intraspécifique. Par conséquent, les autopolyploïdes présentent des génomes dupliqués, et donc des chromosomes d'une très grande identité ou homologie : les chromosomes sont dits homologues.

L'allopolyploïdie est la duplication par hybridation interspécifique ou intergénérique. Elle résulte de l'association de deux ou plusieurs génomes d'espèces distinctes mais proches phylogénétiquement (Soltis and Soltis, 2000; Wendel, 2000; Wendel and Doyle, 2005). Les chromosomes sont dits homéologues : ils présentent moins d'homologie que dans un autopolyploïde, ils sont similaires, mais les gènes ne sont pas toujours conservés et leur ordre peut être différent.

La polyploïdie ne consiste pas en la simple addition des génomes : les polyploïdes subissent des changements génétiques, fonctionnelles et épigénétiques complexes (Ozkan et al., 2006; Stupar et al., 2007; Parisod et al., 2010), tels que des échanges entre chromosomes homéologues (Chalhoub et al., 2014 ; IWGSC, 2014), des pertes de gènes ou des modifications épigénétiques (Liu and Wendel, 2003; Adams and Wendel, 2005b ; Chen, 2007 ; Doyle et al., 2008 ; Leitch and Leitch, 2008 ; Jackson and Chen, 2010 ; Brenchley et al., 2012 ; IWGSC, 2014).

La polyploïdie induit également une reprogrammation importante de l'expression des gènes qui se retrouvent, majoritairement, dupliqués (Adams and Wendel, 2005; Gaeta et al., 2007; Ha et al., 2009b; Rapp et al., 2009; Chague et al., 2010; Flagel and Wendel, 2010; Chelaifa et al., 2013).

1.1.2. Mécanismes de formation des polyploïdes

Il existe deux types de divisions cellulaires chez les eucaryotes : la mitose, qui concerne les cellules somatiques et assure la naissance de cellules identiques à la cellule mère lors de la multiplication asexuée (x=2n) et la méiose, qui aboutit à la production de gamètes (cellules monoploïdes ou haploïdes, x=n) pour la reproduction. La formation de polyploïdes, par doublement du stock chromosomique, peut emprunter l'une ou l'autre de ces deux voies: la voie somatique et la voie méiotique (formation puis fusion de gamètes non réduits).

1.1.2.1. Formation par voie somatique

Lors de la mitose, la cellule somatique subit une division cellulaire suite à la réplication des chromosomes. Lorsque la réplication des chromosomes n'est pas suivie par la division cellulaire, la cellule somatique devient polyploïde. Le doublement somatique peut se produire dans des cellules tissulaires, dans un zygote ou dans un jeune embryon, générant, pour ces deux derniers cas un individu polyploïde (Ramsey and Schemske, 1998). Cette voie de doublement chromosomique est utilisée, en laboratoire, *via* des agents chimiques bloquant la division cellulaire tel que la colchicine pour obtenir des autopolyploïdes ou des allopolyploïdes synthétiques après une hybridation interspécifique.

Toutefois, les espèces polyploïdes empruntent, plus naturellement, la voie des gamètes non-réduites (Harlan and deWet, 1975; Bretagnolle and Thompson, 1995; Ramsey and Schemske, 1998), impliquant la méiose (cf ci-dessous).

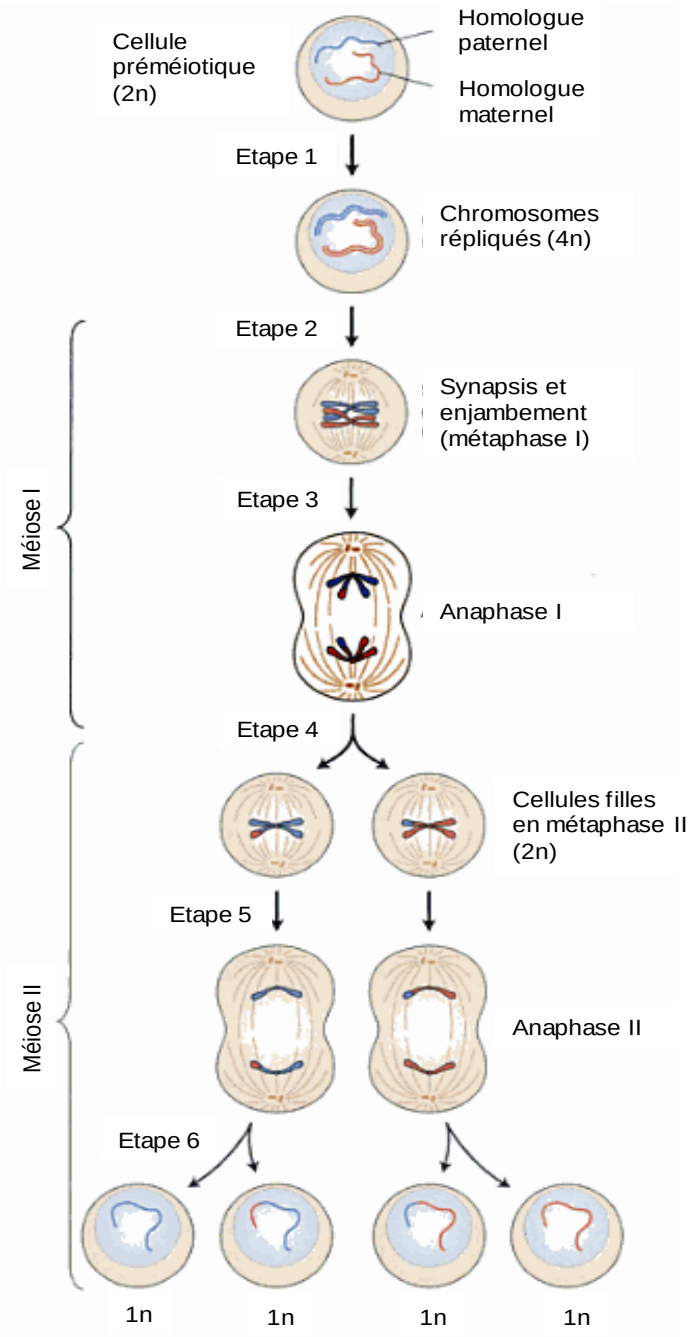


Figure 3: La méiose.
D'après (Lodish et al., 2005). Les cellules préméiotiques possèdent deux copies de chaque chromosome (2n), l'un paternel, l'autre maternel. Etape 1 : avant la première division méiotique, tous les chromosomes sont répliqués durant la phase S, aboutissant à une cellule 4n. Etape 2 : pendant que les chromosomes se condensent durant la première phase méiotique, les chromosomes homologues répliqués s'apparient par au moins un enjambement. Cet appariement est appelé synapsis. Etape 3 : A l'anaphase I, les chromosomes homologues à 2 chromatides migrent vers les pôles opposés du fuseau. Etape 4 : La cytocinèse fournit deux cellules filles (2n), qui entrent dans la méiose II sans passer par une réplication de l'ADN. Etape 5 et 6 : La ségrégation des chromatides vers les pôles opposés du fuseau durant la seconde anaphase méiotique suivie d'une cytocinèse génère des cellules germinales haploïdes (1n), contenant une copie de chaque chromosome.

1.1.2.2. Formation par voie méiotique

La méiose produit des cellules germinales haploïdes (ovules et pollens chez les plantes) qui peuvent fusionner pour générer des zygotes diploïdes.

A la méiose, une seule phase de réplication est suivie de deux cycles de divisions cellulaires, la méiose I et la méiose II. L'enjambement ou « crossing-over » des chromatides à la première métaphase méiotique permet la recombinaison entre les chromosomes parentaux. A la méiose I, les chromosomes homologues, à deux chromatides, se positionnent sur le plan métaphasique, puis chaque chromosome (à deux chromatides) ségrége vers les pôles opposés du fuseau. Pour la méiose II, chaque chromatide de chaque chromosome ségrégent vers les pôles opposés, générant ainsi 4 cellules haploïdes, dites gamètes réduits (Fig. 3).

Il est admis ainsi que les allopolyploïdes peuvent êtes générés dans la nature selon deux mécanismes: par une hybridation interspécifique entre deux gamètes non-réduits, générant directement un individu allopolyploïde ou par une hybridation de deux gamètes réduits d'espèces différentes, générant des hybrides interspécifiques haploïdes, qui peuvent suite à un doublement spontané du génome, générer une plante allopolyploïde.

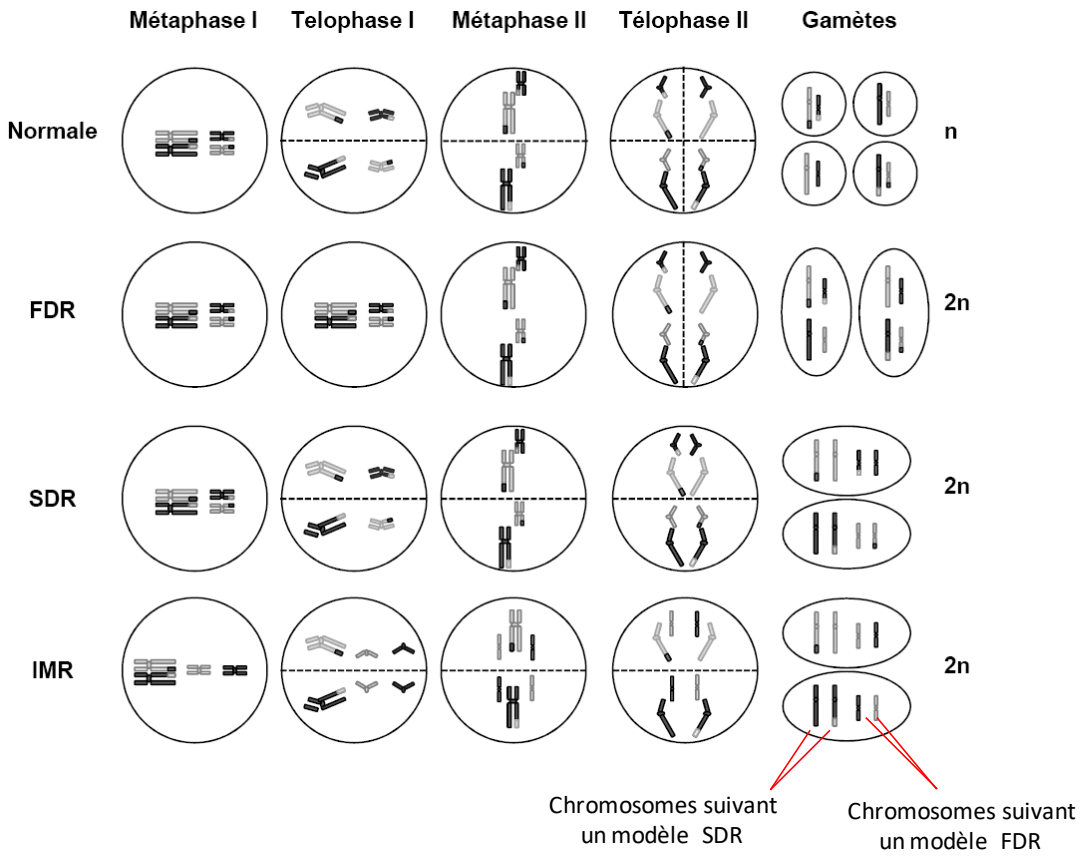


Figure 4: Formation de gamètes non réduits (2n) de types FDR (First Division Restitution), SDR (Seconde Division Restitution), et IMR (Indeterminate Meiotic Restitution).

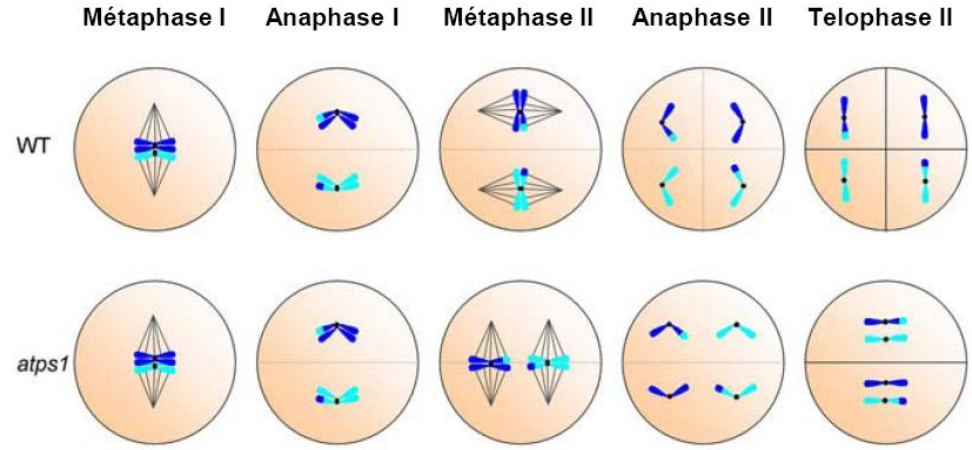


Figure 5: Formation de gamètes non réduits (2n) suite à une orientation parallèle des fibres des fuseaux chez les mutants atps1 d'Arabidopsis.

D'après (Andreuzza and Siddiqui, 2008 ; d'Erfurth et al., 2008).

1.1.2.2.1. Formation des gamètes non-réduits suivie d'une hybridation interspécifique

Une erreur à la méiose peut induire la formation de gamètes diploïdes non réduits. Ces « gamètes 2n » contiennent le nombre total de chromosome somatique. Il existe différents types de gamètes non-réduits selon leurs voies ou leurs stades de formation (Bretagnolle and Thompson, 1995 ; Ramsey and Schemske, 1998) (Fig. 4):

- les gamètes FDR (First Division Restitution) sont issus de la non-disjonction des chromosomes homologues à la première division méiotique. La méiose aboutit alors à des gamètes 2n hétérozygotes aux crossing-overs près et toutes identiques ;

- les gamètes SDR (Second Division Restitution) proviennent de la non-disjonction des chromatides sœurs à la deuxième division méiotique. Il y a ainsi formation de gamètes 2n homozygotes aux crossing-overs près et toutes différentes ;

- les gamètes IMR (Indeterminate Meiotic Restitution) sont des gamètes 2n pour lesquelles certains chromosomes ont subi une FDR et d'autres une SDR ;

- les gamètes issus d'une absence totale de division méiotique (FDR suivie d'une SDR) possèdent les deux paires de chromosomes homologues à deux chromatides.

Les gamètes non réduits peuvent également être formées suite à une orientation anormale des fuseaux lors de la méiose. Ainsi, la formation de fuseaux parallèles lors de la métaphase II peut générer des gamètes non réduits de type FDR. Ceci a été décrit pour les mutants ps1 (parallel spindle 1) de pomme de terre (Andreuzza and Siddiqui, 2008), ou encore le gène atps1 (Arabidopsis thaliana parallel spindle 1) qui entraine la formation de fuseaux en orientation parallèle à la métaphase II aboutissant à la formation de gamètes non-réduits (2n) (d'Erfurth et al., 2008) (Fig. 5). Plusieurs travaux (d'Erfurth et al., 2008 ; d'Erfurth et al., 2009 ; Erilova et al., 2009; d'Erfurth et al., 2010 ; Brownfield and Kohler, 2011) ont permis d'identifier des gènes, dont les mutations induisent la formation de gamètes non-réduites.

1.1.2.2.2. Formation par hybridation entre gamètes réduits suivie d'un doublement du génome

Des hybrides interspécifiques peuvent souvent se former par hybridation entre des gamètes normaux réduits et haploïdes. Ces hybrides interspécifiques reçoivent normalement un ensemble de chromosomes haploïdes provenant de chaque espèce parentale et sont stériles car il n'y a pas d'appariement et/ou crossing-over entre chromosomes pour que la méiose ait lieu. Le doublement des chromosomes permet à l'hybride interspécifique de devenir un allopolyploïde et rétablit donc l'appariement des chromosomes homologues qui se retrouvent en deux copies à la méiose. L'allopolyploïde devient fertile avec un cycle de reproduction normalement stable.

L'hybridation intra ou interspécifique entre deux gamètes non réduits ou un gamète non-réduit et un réduit aboutit à la formation d'un polyploïde. L'union de gamètes réduits (n) et non-réduits (2n) peut aussi aboutir à la formation de polyploïdes avec un nombre impair de jeux de chromosomes (triploïde, pentaploïde…). Ces polyploïdes sont souvent stériles, car leur méiose ne peut se dérouler correctement du fait d'un problème d'appariement des chromosomes uniques. Cependant, un doublement somatique dans ces individus ou la fusion de leurs gamètes non réduits (sans division) entre eux ou avec des gamètes parentaux (n) peut leurs permettre de former des polyploïdes stables et fertiles.

Au cours de mes travaux de thèse, je me suis intéressée au blé tendre, qui est un modèle allohexaploïde issu de deux évènements de polyploïdisations relativement récents. Le premier événement de polyploïdisation a eu lieu il y a environ 500 000 ans, entre deux espèces diploïdes *Triticum urartu* (porteuse du génome AA, 2n=2x=14) et une espèce diploïde *Aegilops speltoides* (porteuse du génome BB, 2n=2x=14) pour former une espèce tétraploïde *Triticum turgidum* (porteuse du génome BBAA, 2n=4x=28) également nommé « blé dur ». Le deuxième événement, plus récent, a eu lieu entre ce blé tétraploïde et un blé diploïde *Aegliops tauschii* (porteur du génome DD, 2n=2x=14) (Feldman et al., 1995), que je détaille dans la partie 1.2. En laboratoire, ce croisement donne un hybride F1 triploïde stérile qui, une fois le génome doublé (spontanément ou à l'aide de la colchicine), aboutit à l'hexaploïde fertile.

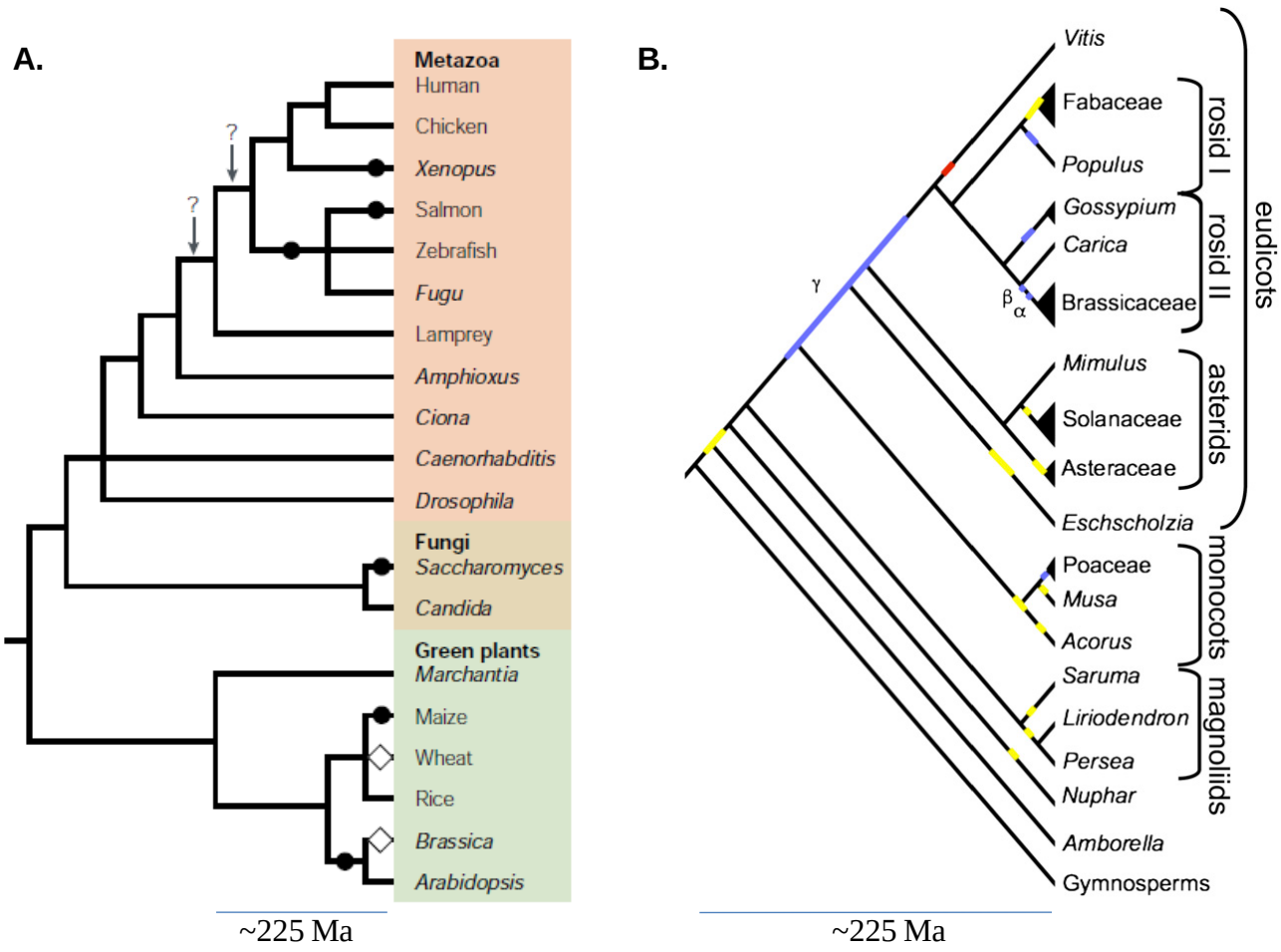


Figure 6: Evènements de polyploïdisation au cours de l'évolution des eucaryotes et des angiospermes.

(A). Polyploïdisation chez les eucaryotes d'après (Wolfe, 2001). Les points d'interrogation marquent les emplacements possibles de 2 cycles de duplication globale. Le cercle noir indique un évènement de polyploïdisation (allo- ou auto-, le diamant indique 2 évènements de polyploïdisation. (B). Polyploïdisation chez les angiospermes d'après (Soltis and Soltis, 2009). Les barres bleues et jaunes indiquent des duplications entières de génomes montrées par l'analyse de la séquence de génome complet ou d'EST. La barre rouge indique une duplication alternative de la vigne (Velasco et al. 2007). Le moment précis de la duplication γ est incertain, notamment pour savoir si elle est commune ou non aux plantes monocotylédones.

1.1.3. Importance de la polyploïdie

La polyploïdie est récurrente et ubiquitaire (Fig. 6), elle joue un rôle majeur dans l'évolution des eucaryotes, en particulier les angiospermes (Wendel, 2000 ; Wolfe, 2001 ; Blanc and Wolfe, 2004 ; Schlueter et al., 2004 ; Adams and Wendel, 2005b ; Cui et al., 2006 ; Jaillon et al., 2007 ; Paterson and al., 2009; Soltis and Soltis, 2009 ; Jiao et al., 2011). Tous les angiospermes ont subi au moins un événement de polyploïdisation (Van de Peer et al., 2009b ; Van de Peer et al., 2009a ; Project., 2013).

Historiquement, la polyploïdie a été découverte il y a plus d'un siècle, depuis les travaux de DeVries (Lutz, 1907 ; Gates, 1908) sur des polyploïdes relativement récents dont on pouvait distinguer les chromosomes dupliqués grace à un faible remaniement de ceux-ci. De nombreux travaux, basés tout d'abord sur les analyses cytologiques, (Muntzing, 1936; Darlington, 1937; Stebbins, 1950; Grant, 1981; Goldblatt, 1980; Soltis and Soltis, 2009), puis les travaux de séquençage d'EST (Expressed Sequenced Tag) ainsi que le séquençage complet des génomes diploïdes tels que le riz (Project., 2005), le maïs (Wei et al., 2009), *Arabidopsis thaliana* (Initiative, 2000), le peuplier (Tuskan et al., 2006), la vigne (Jaillon et al., 2007 ; Velasco et al., 2007), le papayer (Ming et al., 2008), la tomate (Consortium., 2012) et plus récemment l'Amborella (Project., 2013) ont montré que la polyploïdie est un mécanisme fréquent et récurent chez les plantes à fleurs (angiospermes), dont toutes les espèces ont subi des cycles répétés de duplication de génome (Wendel, 2000; Cui et al., 2006; Jiao et al., 2011 ; Project., 2013).

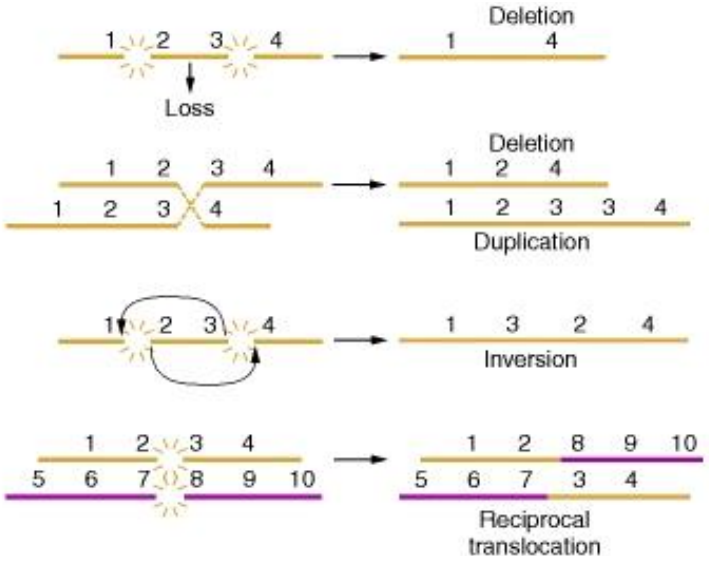
Contrairement aux végétaux, l'existence d'espèces polyploïdes est plus rare chez les animaux, car la polyploïdie conduit souvent à la mortalité et la stérilité.

1.1.4. Les conséquences de la polyploïdie

La polyploïdie joue un rôle majeur dans l'évolution des plantes, et induit des changements génomiques, morphologiques, physiologique et écologique sur une ou plusieurs générations, permettant aux polyploïdes de mieux s'adapter à des environnements plus extrêmes que ceux tolérés par leurs espèces parentales. Un niveau élevé de tolérance à la sécheresse et une plus grande résistance à des parasites ont permis aux polyploïdes de mieux survivre à des environnements défavorables (hautes altitudes, climat froid) comparés à leurs parents diploïdes (Chen, 2007). La polyploïdie est un facteur important dans la diversification et la prolifération des espèces, par la « pré-adaptation ». En effet, des polyploïdes sont prédisposés à des nouvelles conditions, et présentent des meilleurs taux d'adaptation et de survie dès les premières générations. Cette augmentation de la diversité génétique leurs permet au cours de l'évolution d'augmenter leur capacité de prolifération (te Beest et al., 2012). Au niveau physiologique, l'allopolyploïdie est souvent associée à une meilleure vigueur des plantes et une meilleure adaptation en présence de stress environnementaux (Comai, 2005 ; Chen, 2007 ; Soltis et al., 2008).

Les données à l'échelle du génome ont montré que la polyploïdie joue un rôle clé dans le modelage des génomes des plantes. Barbara McClintock (McClintock, 1984) formule le terme de « choc génomique », qu'elle définie comme une réponse à l'hybridation et autres événements de stress. Elle prédit le déroulement de changements structuraux, tels que des translocations, inversions, délétions et duplications dans le génome. Ces deux dernières décennies, les analyses génomiques, épigénétiques et cytogénétiques ont validé ces prédictions dans plusieurs systèmes allopolyploïdes (Hufton and Panopoulou, 2009 ; Jackson and Chen, 2010 ; Mayfield et al., 2011 ; De Smet and Van de Peer, 2012 ; Hernandez et al., 2012; Heslop-Harrison, 2012 ; Soltis and Soltis, 2012 ; Ma et al., 2013; IWGSC, 2014).

A. Réarrangement chromosomique par coupure suivie d'une ligation



B. Réarrangement chromosomique par crossing-over entre ADN repété

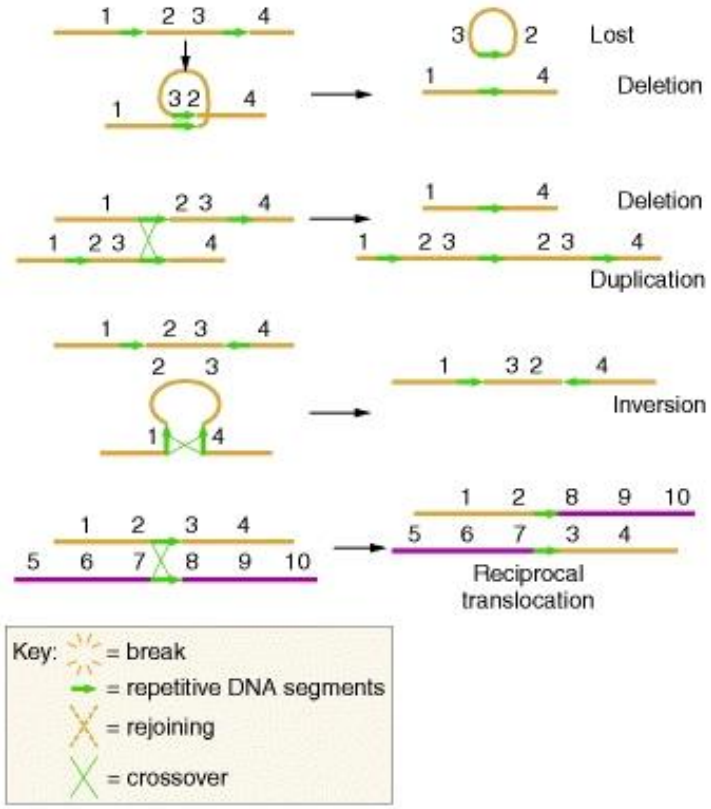


Figure 7: Origine des réarrangements chromosomiques.

D'après (Griffiths et al., 1999).

1.1.4.1. Les changements structuraux des génomes polyploïdes

Les réarrangements chromosomiques regroupent les délétions[1], les duplications, les inversions[2], les translocations d'ADN ainsi que les échanges entre chromosomes homéologues. Chacun de ces événements peut être causé par une cassure de l'ADN à deux endroits différents, suivis d'une ligation croisée des extrémités cassées induisant un nouvel arrangement des gènes, différent de l'ordre initial (Griffiths et al., 1999), ou par crossing-over (Fig. 7).

Au niveau du segment de chromosome, les échanges entre homéologues (HE), sont appelés également translocations, transpositions réciproques, ou non-réciproques des homéologues (HNRT), ainsi que conversions géniques. Ils sont caractérisés par la perte de régions génomiques plus au moins grandes (de quelques nucléotides à des chromosomes entiers) remplacées par une copie dupliquée de la région homéologue correspondant (Chalhoub et al., 2014).

Il existe deux types de réarrangements : équilibrés ou non-équilibrés. Les réarrangements équilibrés ne s'accompagnent pas généralement de perte ou de gain de matériel génétique (inversion et translocation). Ces réarrangements sont observés chez la levure, les vertébrés, dans les cancers chez les humains et plus fréquemment chez les plantes (Hufton and Panopoulou, 2009). Les remaniements non-équilibrés sont d'autant plus graves que la perte ou le gain de matériel est important (délétion, duplication d'un fragment de chromosome). L'aneuploïdie (la perte ou le gain d'un ou plusieurs chromosomes entiers) est une forme de réarrangement déséquilibré (que je détaille plus loin dans ce même paragraphe).

Les réarrangements chromosomiques sont mis en évidence par comparaison avec les espèces parentales par des méthodes de cytogénétique et par l'utilisation de marqueurs moléculaires. Chez les anciens (paléo) polyploïdes, ces réarrangements génomiques sont mis en évidence par comparaisons de séquences. Les segments de synténie conservée (ordre des gènes

[1] Une délétion résulte d'une cassure chromosomique avec perte du segment distal (délétion terminale) ou de deux cassures sur un même bras avec perte du segment intercalaire (délétion interstitielle ou intercalaire).

[2] Une inversion résulte de deux cassures sur un même chromosome suivies de recollement après inversion du segment intermédiaire. Elle est paracentrique si les points de cassure sont localisés sur un même bras chromosomique, et péricentrique si les points de cassure sont localisés de part et d'autre du centromère. Cette anomalie entraine des difficultés d'appariement à la méiose.

conservé) peuvent être identifiés entre les espèces apparentées, la perte de synténie pouvant traduire des réarrangements du génome (Hufton and Panopoulou, 2009). Différents types de changements chromosomiques, prédits par McClintock (1984), tels que des réarrangements intergénomiques, des délétions, des aneuploïdies ont été observés dès les premières générations du colza synthétique *Brassica napus* (AACC) (un allotétraploïde issu du croisement entre *B. oleracea* (CC) et *B. rapa* (AA) (Gaeta et al., 2007 ; Xiong et al., 2011 ; Chalhoub et al., 2014), des polyploïdes naturels *Tragopogon. micellus* et *T. mirus* (Lim et al., 2008 ; Chester et al., 2012), et *Arabidopsis suecica* (Wright et al., 2009 ; Matsushita et al., 2012). Néanmoins, peu de changements génomiques ont été observés chez les allopolyploïdes récents tel que la spartine (Baumel et al., 2002 ; Salmon et al., 2005).

Chez le colza, les réarrangements génomiques sont plus fréquents dans les générations S5 suivant la polyploïdisation (Gaeta et al., 2007). Toutefois, le séquençage récent du génome du colza naturel, constituant le premier polyploïde récent séquencé, montre des échanges de séquences homologues dus aux mécanismes de crossing-overs et non crossing-overs fréquents mais une perte de gène quasi infine (Chalhoub et al., 2014).

Des échanges entre chromosomes ou séquences homéologues (HEs) sont observés chez les allopolyploïdes naturels, notamment chez le tabac *Nicotiana tabaccum* (SSTT, 2n=4x=48) (Leitch et Bennett, 1997), le colza (*Brassica napus* : (Udall et al., 2005 ; Chalhoub et al., 2014)) et le blé (HNRT : A4-D7 ou B4-D7, (Feldman et al., 2012; Marone et al., 2012)Badaeva et al., 2007).

Il a été suggéré que des remaniements structuraux concerneraient principalement les séquences non codantes, notamment chez le blé allopolyploïde (Feldman et al., 1997 ; Liu, 1998 ; Liu et al., 1998 ; Ozkan et al., 2001 ; Shaked et al., 2001 ; Han et al., 2005 ; Khasdan, 2010) même si des perte de gènes ont aussi été mises en évidence (pour le blé :(Kashkush, 2002, Pontes et al., 2004), pour les Brassica : (Chalhoub et al., 2014); pour les gènes codant les ARN ribosomiques (rRNA) du fraisier (Liu and Davis, 2011), et également Tragopogon (Malinska et al., 2010)). Des réarrangements au niveau de loci sont aussi présents, chez les allopolyploïdes naturels blé, par exemple, pour le locus *Ph1* (Griffiths et al., 2006 ; Al-Kaff, 2008), et le locus *Ha* (Chantret et al., 2005 ; Li et al., 2008b ; Ragupathy and Cloutier, 2008).

Plusieurs études ont montré le caractère uniparental de l'élimination de chromosomes chez des hybrides interspécifiques issus de croisement entre espèces plus ou moins apparentées ((Kasha, 1970 ; Subrahmanyam, 1977 ; Finch, 1983 ; Zenkteler, 1984 ; Laurie, 1986 ; 1988 ; Rines, 1990 ; Chen, 1991 ; Matzk, 1994 ; 1996 ; 1997) cités dans (Gernand et al., 2005)).

Les travaux portant sur les remaniements structuraux, induits par la polyploïdie (et l'hybridation interspécifique), mènent toutefois à des résultats contrastés quant à la fréquence et à la durée des remaniements structuraux. D'importants remaniements structuraux ont été observés chez les allopolyploïdes synthétiques (et naturels) de blé (Feldman et al., 1997 ; Liu et al., 1998 ; Ozkan et al., 2001 ; Han et al., 2005 ; Hernandez et al., 2012; Ma et al., 2013 ; IWGSC, 2014), alors que des études faites dans notre laboratoire sur les blés hexaploïdes synthétiques ont conclu à un faible taux de remaniement (Mestiri et al., 2010).

Bien que Liu et al (Liu et al., 2001) n'aient montré que peu ou pas de changements génomiques structuraux chez le coton allopolyploïde, les travaux de séquençage de Page et al. (Page et al., 2013a) ont révélé de nombreux changements structuraux tels que des délétions ou des HEs entre chromosomes homéologues dans le coton allotétraploïde.

L'aneuploïdie est plus tolérée chez les polyploïdes. L'obtention de plantes aneuploïdes est, dans la plupart des cas, due à des accidents lors de la ségrégation des chromosomes (ou chromatides) pendant la méiose ou la mitose. L'aneuploïdie se produit spontanément. La fréquence des gamètes aneuploïdes varient entre les espèces et selon le type de polyploïdie (Comai, 2005). Elle est la cause d'anomalies du développement et d'altérations phénotypiques dans toutes les espèces. Elle induit des impacts phénotypiques plus importants que la polyploïdie (Birchler and Veitia, 2007). Les gènes régulateurs maintiennent un équilibre stœchiométrique, qui est perturbé chez la plante aneuploïde (Birchler and Veitia, 2007). L'ensemble du déséquilibre stœchiométrique des protéines est à l'origine de perturbations cellulaires, d'anomalies du développement et de phénotypes observés associés à l'aneuploïdie (Torres et al., 2008).

L'aneuploïdie est très étudiée, chez l'homme, notamment pour le syndrome de Down (trisomie 21), le syndrome de Klinefeleter (XXY), le syndrome de Turner (X0), ou encore au niveau des cellules cancéreuses chez les mammifères (Lengauer et al., 1998 ; Pihan and Doxsey, 2003). Chez la levure, *Saccharomyces cerevisiae*, l'aneuploïdie est bien tolérée (Parry and Cox, 1970). L'aneuploïdie mitotique a été montrée chez l'hexaploïde *Senecio cambrensis* (Ingram and Noltie, 1995), suggérant que le niveau élevé de ploïdie permet à une meilleure tolérance à l'aneuploïdie (Coyne J.A and H.A, 2004).

Les remaniements structuraux et l'aneuploïdie peuvent être tolérées jusqu'à un certain seuil, au-delà duquel ils mènent à l'extinction (Xiong et al., 2011 ; Matsushita et al., 2012). Cette forme de variabilité cytogénétique entre les individus peut-être considérée comme une réserve disponible pour la diversité, la sélection ou sinon comme une forme d'instabilité génomique tolérée jusqu'à un certain seuil (Madlung and Wendel, 2013).

1.1.4.2. La réorganisation de l'expression des gènes

Les variations phénotypiques observées chez les allopolyploïdes par rapport à leur progéniteurs, suggèrent des modifications dans l'expression et la régulation des gènes dont la plupart se retrouvent en copies dupliquées chez le polyploïde.

Le choc transcriptomique suite à l'allopolyploïdie a été observé chez *Senecio* (Hegarty et al., 2005 ; Hegarty et al., 2006), *Tragopogon* (Buggs et al., 2011b), le blé (He et al., 2003 ; Chague et al., 2010), le colza (Cui et al., 2013 ; Chalhoub et al., 2014) et chez les hybrides F1 de la spartine (Chelaifa et al., 2010a ; Chelaifa et al., 2010b). L'expression des gènes dans le polyploïde pourrait être régulée post transcription, par un épissage alternatif, qui aboutiraient à différentes protéines à partir d'un même mRNA (chez *Arabidopsis* (Madlung et al., 2005 ; Zhang et al., 2011), *Brassica* (Zhou et al., 2011), *Capsella bursapastoris* (Slotte et al., 2009)).
Les changements de conditions environnementales induiraient aussi des changements d'expression dans les polyploïdes par rapport aux progéniteurs (Bardil et al., 2011 ; Dong and Adams, 2011).

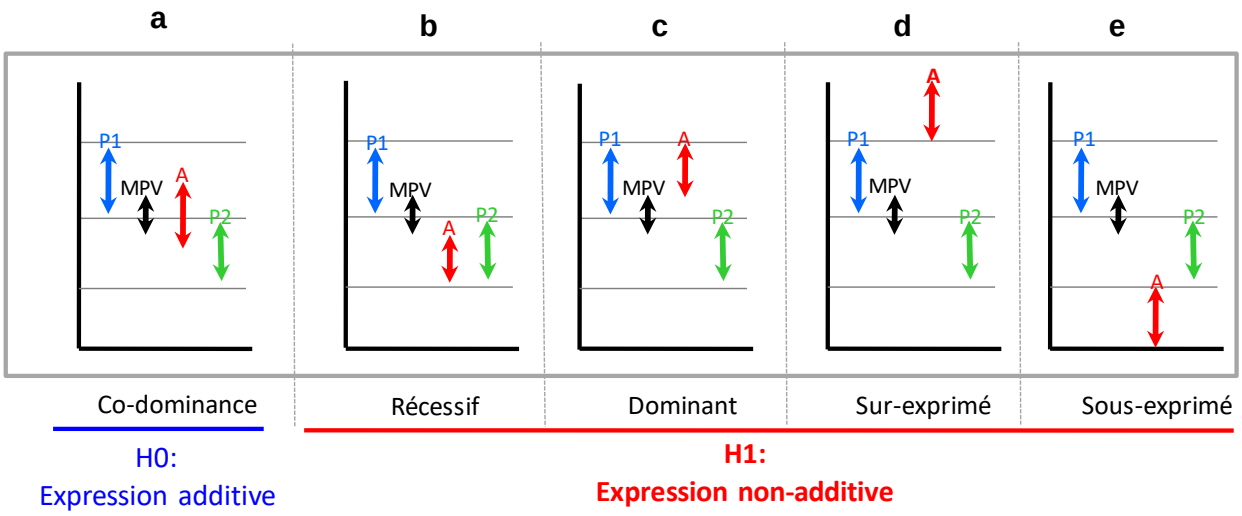


Figure 8: Modèle d'interprétation de la reprogrammation de l'expression des gènes dans l'allopolyploïde.

Abréviations : A, allopolyploïde ; P1, parent 1 ; P2, parent 2 ; MPV, Mid-Parent Value. Sous l'hypothèse d'expression additive, le niveau d'expression du gène dans l'allopolyploïde révèle une co-dominance de l'expression des parents. Sous l'hypothèse de non-additivité : le niveau d'expression dans l'allopolyploïde peut-être (8b) égal à celui du parent récessif, (8c) égal à celui du parent dominant, (8d) être sur-exprimé par rapport à l'expression des parents, (8e) encore sous-exprimé.

L'estimation de la reprogrammation de l'expression des gènes dans les polyploïdes a évolué en fonction des progrès techniques et génomiques. Les premières techniques telles que de la PCR, les microarrays ne permettaient pas en général de distinguer l'expression des copies de gènes dupliqués, actuellement de nouvelles techniques dont le RNA-Seq issus des nouvelles générations de séquençage permettent un revirement de situation.

1.1.4.2.1. Estimation des variations de l'expression des gènes dans les polyploïdes

1.1.4.2.1.1. Les techniques ne distinguant pas les copies dupliquées des gènes

Pour caractériser la reprogrammation de l'expression des gènes dans les différents modèles allopolyploïdes, une première méthode consiste à comparer, pour un même gène, la différence des niveaux d'expression entre l'allopolyploïde et ses progéniteurs.

Le modèle d'interprétation, de la reprogrammation de l'expression des gènes chez les blés allopolyploïdes que j'ai utilisé au cours de ma thèse, repose sur deux hypothèses alternatives (Fig. 8):

- l'hypothèse nulle (H_0) selon laquelle le niveau d'expression du gène est similaire à celui de la moyenne de ses parents dite MPV (Mid Parent Value). L'expression est dite alors additive et dans ce cas on considère qu'il n'y a pas d'effet de l'allopolyploïdie, sur l'expression des gènes (Fig. 8a).

- l'hypothèse alternative (H_1) consiste à tester que l'allopolyploïdie induit une reprogrammation de l'expression. L'expression est non-additive (Chen, 2010). Dans ce cas, le niveau de l'expression du gène dans l'allopolyploïde diffère de la MPV. Le niveau d'expression du gène, dans l'allopolyploïde, peut être inférieur à la MPV (Fig. 8b), ou supérieur à la MPV (Fig. 8c). Le niveau peut être sur-exprimé (Fig. 8d) dans le cas où il est simultanément supérieur à la MPV et ses deux progéniteurs et sous-exprimée dans le cas inverse.

Dans notre laboratoire, la MPV est mesurée par hybridation d'un mélange équimolaire des ARN parentaux (Wang et al., 2006b ; Gaeta et al., 2009 ; Chague et al., 2010 ; Chelaifa et al., 2013).

D'autres études ont considéré une MPV *in silico*, en moyennant les valeurs d'expression observées dans les progéniteurs (Pumphrey et al., 2009 ; Rapp et al., 2009). Les deux types d'estimation de la MPV ont été comparés (Chague et al., 2010) et montrent que la MPV *in silico* explique seulement 89.67–91.57% de la variance de son estimation *in vitro*. L'utilisation de la MPV *in silico* néglige les biais techniques et les interactions entre homéologues d'où l'intérêt de la précision apporté par la MPV *in vitro* dans mon étude.

L'écart à l'additivité varie selon les espèces: ~5% des gènes chez les allotétraploïdes synthétiques d'*Arabidopsis* (Wang et al., 2006b), 1 à 6% chez le coton (Adams and Wendel, 2004 ; Rapp et al., 2009) et ~5% chez *Senecio* (Hegarty et al., 2005). Chez le blé, les études d'expression de gènes montrent qu'il y a entre 1 à 18% de gènes avec un profil non-additif, selon la méthode d'étude et le génotype de blé allohexaploïde utilisé (Pumphrey et al., 2009 ; Akhunova et al., 2010 ; Chague et al., 2010 ; Qi et al., 2012 ; Chelaifa et al., 2013).
Par l'utilisation des puces Affymetrix GeneChip® blé (55 049 transcrits), les études effectuées dans notre laboratoire montrent 1-7% de gènes non-additifs chez les allohexaploïdes synthétiques (Chague et al., 2010 ; Chelaifa et al., 2013) suggérant que la reprogrammation de l'expression des gènes est rapidement établie lors de l'allopolyploïdie et hautement conservée à travers les générations, comme le montre la comparaison faite entre les allohexaploïdes synthétiques et naturels de blé.

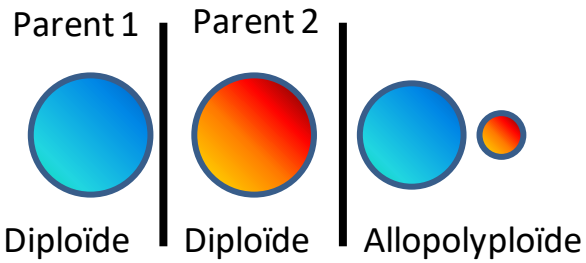


Figure 9: Biais d'expression d'homéologue.

D'après (Grover et al., 2012). Les parents 1 et 2 sont diploïdes. Pour un gène donné, l'expression de leur homéologue est à un niveau similaire. Dans l'allopolyploïde, l'expression des homéoallèles parentaux est modifiée, dans ce cas, l'homéologue du parent 1 est sur-exprimé par rapport à celui du parent 2. C'est un biais d'expression d'homéologue vers le parent 1.

1.1.4.2.1.2. Les outils distinguant l'expression des copies homéologues

Bien entendu, l'objectif ultime est de comparer l'expression de chacune des copies homéologues dupliquées ainsi que d'estimer leur contribution à l'expression globale des gènes. Ceci commençait à devenir possible avec le développement de sondes microarray homéologues spécifiques et s'est généralisé avec les outils RNA-Seq (Akhunova et al., 2010 ; Flagel et al., 2008). Au cours de ma thèse, je me suis intéressée à l'utilisation de ces outils sur les blés polyploïdes.

L'objectif des outils homéologues spécifiques est de permettre une distinction de l'expression de chacune des copies des gènes dupliqués. Dans un allopolyploïde, l'expression d'un gène devient 'disséquable' au niveau de chacune de ses copies homéologues, l'expression globale du gène étant la combinaison de l'expression de tous ses homéologues.

La comparaison de l'expression des homéologues, dans le polyploïde, peut révéler un biais provenant de l'expression plus importante d'un homéologue (Grover et al., 2012) par rapport à une autre. Pour un gène donné, le biais de contribution d'homéologue décrit donc une expression relativement majoritaire (et donc dominante) d'un homéologues par rapport à l'autre (ou aux autres) homéologue(s) (Fig. 9). Des biais de contribution d'homéologues ont été montrés pour une proportion de gènes plus au moins importante dans toutes les espèces étudiées : le coton, *Brassica rapa* (Schnable and Freeling, 2011 ; Ilut et al., 2012 ; Tang et al., 2012 ; Yoo et al., 2013) et le colza (*Brassica napus*) (Chalhoub et al., 2014).

On parle de dominance d'un sous génome, si pour une proportion significative de gènes les homéologues d'un des sous-génomes dominent l'expression et donc contribuent plus que ceux de l'autre sous-génome. Au contraire on parle d'équivalence si la dominance des homéologues est distribuée de façon équivalente entre les deux sous-génomes (Wang et al., 2006b ; Rapp et al., 2009 ; Chague et al., 2010 ; Flagel and Wendel, 2010 ; Bardil et al., 2011 ; Chelaifa et al., 2013 ; Chalhoub et al., 2014; Pfeifer et al., 2014).

1.1.4.2.2. Les mécanismes des variations de l'expression

Les régulateurs en cis et trans

Un élément cis-régulateur est une séquence d'ADN qui a un effet régulateur sur la transcription d'un gène. Les facteurs trans-régulateurs sont des protéines, qui se fixent sur ces séquences cis régulatrices, pour moduler la vitesse de la transcription : activateurs (séquences enhancer) ou inhibiteurs (séquences silencer). Les régulations en cis et en trans sont associées à des modifications de la chromatine (Shi et al., 2012).

Plusieurs changements observés chez les polyploïdes font intervenir des facteurs de régulation cis et trans (Stupar and Springer, 2006 ; Chaudhary et al., 2009; Shi et al., 2012). Une étude récente, sur les pétales du coton, suggère que les biais d'expression conservés dans le polyploïde seraient liés aux régulations en cis, qui ont lieu avant la polyploïdisation (Rambani et al., 2014).

Alors que les connaissances sur les bases moléculaires des effets phénotypiques observés chez les allopolyploïdes sont limitées, les divergences des régulations cis et trans entre les génomes parentaux et leurs interactions sont importantes pour la régulation de la transcription dans les hybrides interspécifiques (Landry et al., 2005) et les allopolyploïdes (Chaudhary et al., 2009).

Effet dose

Un changement de dosage a un impact sur le phénotype due à l'altération de l'expression du gène (Birchler et al., 2001 ; Veitia, 2004 ; Birchler et al., 2005 ; Birchler and Veitia, 2007 ; Veitia et al., 2008 ; Birchler et al., 2010 ; Birchler, 2012). Selon l'hypothèse d'un équilibre de dosage (ou « dosage-balance hypothesis »), cet effet serait due à une perturbation de la quantité de copies de gènes et produits de gènes qui fonctionnent selon des rapports stœchiométriques.

Les éléments transposables, détaillés en Annexe 1, sont aussi des intervenants dans la reprogrammation de l'expression des gènes (Belcram, 2014).

1.1.4.3. Les changements épigénétiques dans les polyploïdes

Le terme épigénétique définit des modifications transmissibles qui ne s'accompagnent pas de changements nucléotidiques des séquences (Rapp and Wendel, 2005). Les marques épigénétiques régulent l'accés aux différentes régions du génome et contrôlent ainsi le statut, actif ou inactif, des gènes (statut «on» ou «off»). Les deux intervenants fondamentaux de l'épigénétique sont la méthylation de l'ADN et les petits ARN, leur coordination permet la régulation de l'expression des gènes.

- **Etats de méthylation de l'ADN**

L'ADN est composé de quatre bases : l'adénine (A), la cytosine (C), la guanine (G), et la thymine (T) décrivant le code génétique.

L'ajout du groupement méthyl (CH_3-) à la Cytosine induit la compaction de la chromatine, ce qui empêche l'expression du gène. La méthylation de la cytosine induit donc le statut « off » de la chromatine, et donc l'inactivation du gène. La méthylation des éléments transposables (TEs) semble être un mécanisme de « défense » contre les insertions délétères de transposons (éléments mobiles et autonomes d'ADN qui peuvent se déplacer d'un endroit à un autre sur un même brin d'ADN ou sur un autre brin, détail en Annexe 1) (Yaakov and Kashkush, 2011).

Les marques épigénétiques les plus fréquemment étudiées sont la méthylation des cytosines et les modifications des histones[3]. Les sites ou îlots de reconnaissance d'une séquences d'ADN qui permettent de modifier leurs états de méthylation, sont CGn, CHG et CHH chez les plantes (Cokus et al., 2008), où H peut-être un quelconque nucléotide autre qu'une guanine, alors que chez les mammifères les méthylations sont majoritairement sur les sites CG (Saze et al., 2008).

Des modifications épigénétiques ont lieu dans les polyploïdes comparés à leur parents et induisent de nouvelles variations (Osborn et al., 2003 ; Pfeifer et al., 2014). Les changements de méthylation sont souvent accompagnés de réarrangements génomiques et de changements de l'expression des gènes (Kashkush, 2002 ; Madlung et al., 2005 ; Gaeta et al., 2007).

[3] Protéine régulant le compactage de l'ADN.

Les modifications des histones et des variants d'histone ont également un rôle important sur l'activité transcriptionnelle des gènes chez les plantes (Ni et al., 2009; Deal and Henikoff, 2011).

Des changements de méthylation d'ADN entre les progéniteurs et les allopolyploïdes ont été montrés chez *Arabidopsis* (Madlung et al., 2002 ; Madlung et al., 2005 ; Beaulieu et al., 2009), chez le blé (Shaked et al., 2001 ; Zhao et al., 2011), chez *Brassica* (Lukens et al., 2006 ; Xu et al., 2009 ; Książczyk et al., 2011), chez la spartine (Salmon et al., 2005), chez le pissenlit dandelion (Verhoeven et al., 2010), et chez *Cymbopogon* en réponse à l'autopolyploïdie (Lavania et al., 2012). Chez le colza des modifications de l'état de méthylation ont été montrées à grand échelle (Chalhoub et al., 2014). Chez la spartine, la méthylation des CG montre plus de variations à proximité des transposons, ceci a été mis en évidence lors d'une expérience d'hybridation comparée à une expérience de duplication du génome (Parisod et al., 2009).

Chez le coton (*Gossypium*), les gènes dupliqués par polyploïdie semblent épigénétiquement mis sous silence à la formation du polyploïde (Adams et al., 2003 ; Adams and Wendel, 2005), les suppressions épigénétiques préserveraient les deux copies dupliquées par exemple jusqu'à une spécialisation tissulaire, ainsi l'une s'exprimerait dans un tissus tandis que l'autre s'exprimerait dans un autre tissu ou à un autre stade du développement. A l'échelle du génome, l'épigénétique préserverait des milliers de gènes pour une évaluation ultérieure par sélection naturelle, lors d'un relâchement sous divers environnement (Madlung and Wendel, 2013).

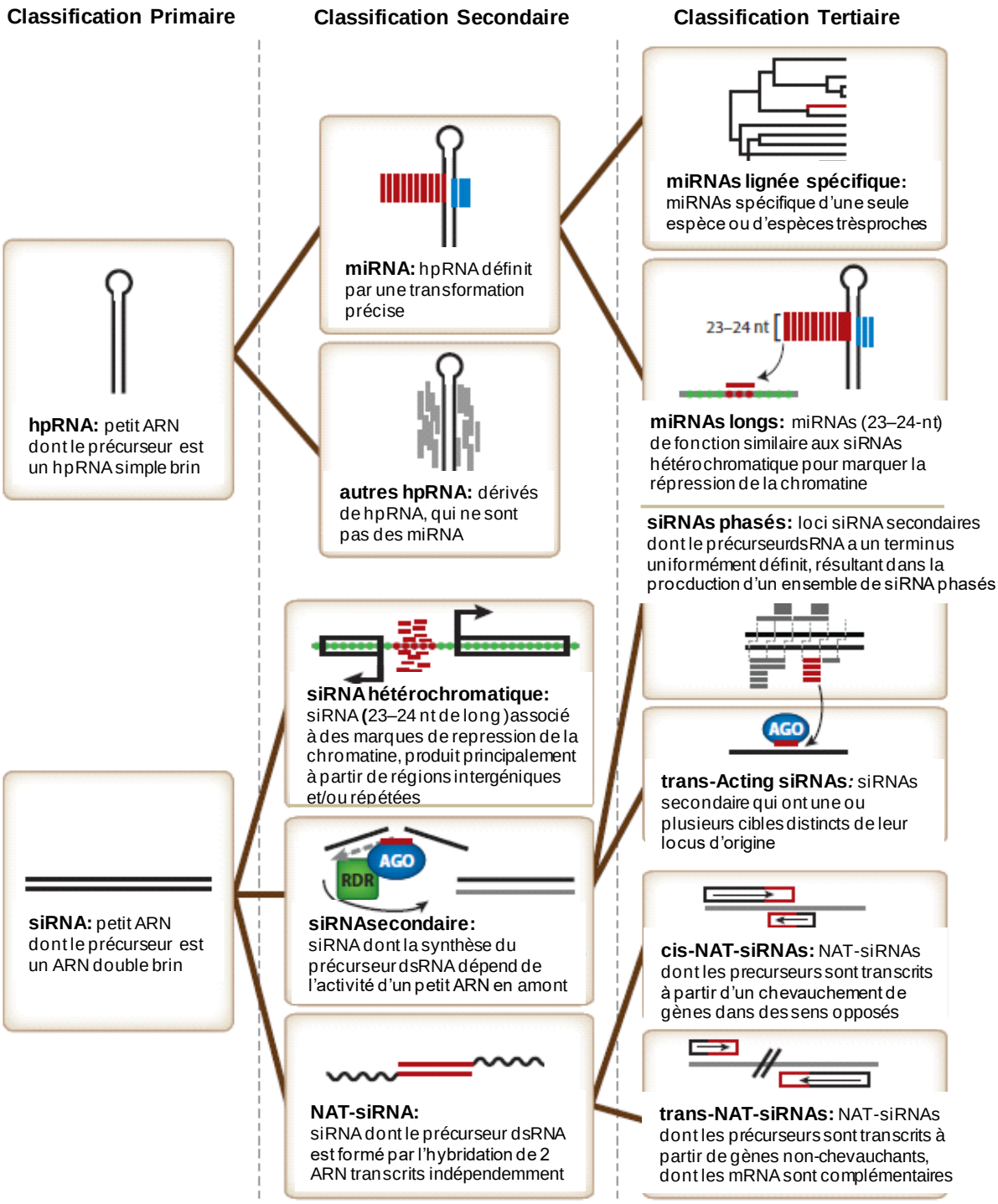


Figure 10: Système de classification hiérarchique des petits ARN endogènes de plantes.

Abréviations: dsRNA, ARN double-brin; hpRNA, hairpin RNA; miRNA, microRNA; NAT-siRNA, natural antisense transcript small interfering RNA; siRNA, small interfering RNA; RDR, RNA-dependent RNA polymerase; AGO, Argonaute. RDRs et AGOs (et Dicer-Like) sont les enzymes essentielles pour la biogénèse et la fonction des petits ARN.

- **Rôle des petits ARN**

Les petits ARN sont transformés à partir de transcrits précurseurs, le produit final affecte les voies de développements, les réponses aux stimuli environnementaux et la stabilité génomique (Chen, 2009). Ils interviennent dans l'épissage des pré-ARNm (snoRNA du splicéosome), la traduction (ARN ribosomaux (rRNA) et ARN de transfert (tRNA)), la régulation des ARN messager (mRNA) par la voie des miRNA et dans le mécanisme de défense contre des ARN exogènes (de virus) par la voie des siRNA (Fig. 10).

Initialement, les petits ARN ont été identifiés comme intervenant dans la régulation des ARNm, par l'ARN interférence (ou RNAi). L'ARN interférence a été découvert dans les années 1990 et a permis de mettre en évidence le mécanisme « Post Transcriptional Gene Silencing » (PTGS, également appelé « RNA silencing ») très répandu chez les plantes. C'est un mécanisme de défense contre les infections virales et les éléments transposables (Waterhouse et al., 2001). Le RNAi consiste en une mise sous silence ou « silencing » génétique post-transcriptionnel induit par l'arrêt de la traduction, ou la dégradation de l'ARNm cible (Ryther, 2004). Il a été découvert par la suite l'effet de l'introduction d'un ARN double-brin (dsRNA). Dans les cellules du nématode *C. elegans*, ce dsRNA induit une diminution significative de la traduction de protéines spécifiques, par hybridation sur leurs mRNA (Fire et al., 1998). Chez les animaux, ce phénomène a été appelé ARN interférence ou siRNA pour «short interference RNA».

Depuis ces premières découvertes, les petits ARN ont été identifiés dans plusieurs processus : la répression de la traduction et le clivage des ARNm, l'inhibition de la transcription *via* la méthylation de l'ADN et l'élimination de l'ADN.

Les petits ARNs sont classés, principalement, en deux catégories (Fig.10) (Axtell, 2013) :

- les miRNA, qui dérivent d'ARN simple brin formant une structure en épingle à cheveux, ce sont des hpRNA (hairpin RNA) ;
- les siRNA, qui ont un précurseur ARN double brin.

D'autres modes de biogenèse de petits ARN existent: chez les animaux, les piwi-associated RNAs (piRNAs) dérivent de la fragmentation d'un précurseur simple brin (Juliano et al., 2011), chez *Caenorhabditis elegans* les ARN 22G sont les produits directs de la transcription (Pak and Fire, 2007). Ces modes de biogenèse de petits ARN n'ont pas été décrits chez les plantes (Axtell, 2013).

Les miRNA

Les miRNA appartiennent à une classe de molécules de 20 à 30 nucléotides de long (Ghildiyal and Zamore, 2009), et se divisent en deux sous-classes (les miRNA lignée spécifique et les long miRNA).

Les miRNA régulent généralement les gènes impliqués dans le développement, par inhibition de la traduction ou par dégradation de l'ARNm cible.

La plupart des miRNA sont transformés par des endonucléases DICER et DICER-LIKE (Ghildiyal and Zamore, 2009), qui transforment les précurseurs d'ARN en petits fragments de nucléotides. Ces fragments sont ensuite liés par des protéines ARGONAUTE pour former le complexe RNA-induced silencing (RISC). Le complexe RISC est guidé par l'hybridation du miRNA complémentaire sur l'ARNm cible pour induire sont clivage.

Les siRNA

Les siRNAs sont des ARN double brin de 24 nucléotides de long. Parmis les siRNA, il y a les siRNA hétérochromatiques, siRNAs secondaires, et les transcrits anti-sens naturel siRNAs (NAT-siRNAs pour « Natural antisense transcript siRNA ») :

- les siRNA hétérochromatiques interviennent dans le modelage de la chromatine ;
- les siRNA secondaires se divisent en siRNA phasés et en trans-acting siRNA (tasiRNA) ;
- les NAT-siRNA peuvent-être cis ou trans (Fig. 10) (Axtell, 2013).

Il existe d'autres classes de petits ARN régulateurs, chez les plantes, tels que les ARN de 30-40 nucléotides de long produits lors d'une infection bactérienne (Katiyar-Agarwal et al., 2007), les petits ARN dérivés de virus ou associés à un pathogène (Ruiz-Ferrer and Voinnet, 2009) et les petits ARN synthétiques (Ossowski et al., 2008; Eamens and Waterhouse, 2011 ; de Felippes et al., 2012).

Les petits ARN, produits pendant l'hybridation interspécifique ou la polyploïdisation, auraient un rôle de tampon lors du choc génomique dans les hybrides interspécifiques et les allopolyploïdes (Ha et al., 2009b).

Les siRNA semblent avoir un rôle de gardien contre les éléments transposables (TEs) pendant le développement des plantes (Levy and Walbot, 1990; Gehring et al., 2009; Hsieh et al., 2009; Slotkin et al., 2009). On notera que les siRNA associés aux régions répétées proviennent des deux espèces progénitrices de l'allopolyploïde chez *Arabidopsis suecica* (Ha et al., 2009b) et le maïs (Barber et al., 2012).

Les miRNA et les tasiRNA régulent l'expression des gènes et le développement chez les plantes et les animaux (Chen, 2009 ; Voinnet, 2009 ; Guan et al., 2014 ; Tian et al., 2014). Les changements phénotypiques et les innovations évolutionnaires possibles dans les allopolyploïdes pourraient être facilités par le changement des modèles ou « patterns » d'expression des miRNA et tasiRNA dans l'allopolyploïde.

Alors que les niveaux d'expression des siRNA sont relativement stables dans les allopolyploïdes, les « patterns » d'expression des miRNA et tasiRNA varient entre l'allopolyploïde et les espèces parentales. L'intégrité du génome est préservée par la compatibilité des siRNA des deux génomes progéniteurs dans l'allopolyploïde, permettant la répression des transposons et des séquences répétées associées aux siRNA (Ha et al., 2009b). L'expression différentielle de miRNA spécifique, dans les allopolyploïdes par rapport à leurs progéniteurs, pourrait mener aux changements évolutifs des nouveaux allopolyploïdes (Gong et al., 2013).

Des travaux menés sur le système *Arabidopsis* (Ha et al., 2009b ; Lackey et al., 2010), suggèrent que la régulation du métabolisme des petits ARN est plus complexe dans les allopolyploïdes comparés à leurs parents diploïdes. Les travaux portant sur le blé, comparant les petits ARN entre espèces diploïdes, tetraploïdes et hexaploïdes ont montré que les niveaux d'expression des siRNA correspondant à des transposons diminuent avec l'augmentation du niveau de ploïdie (Kenan-Eichler et al., 2012), ce qui suggère que l'hybridation et/ou la polyploïdisation pour ces plantes mènent à la déstabilisation du génome (Kenan-Eichler et al., 2012).

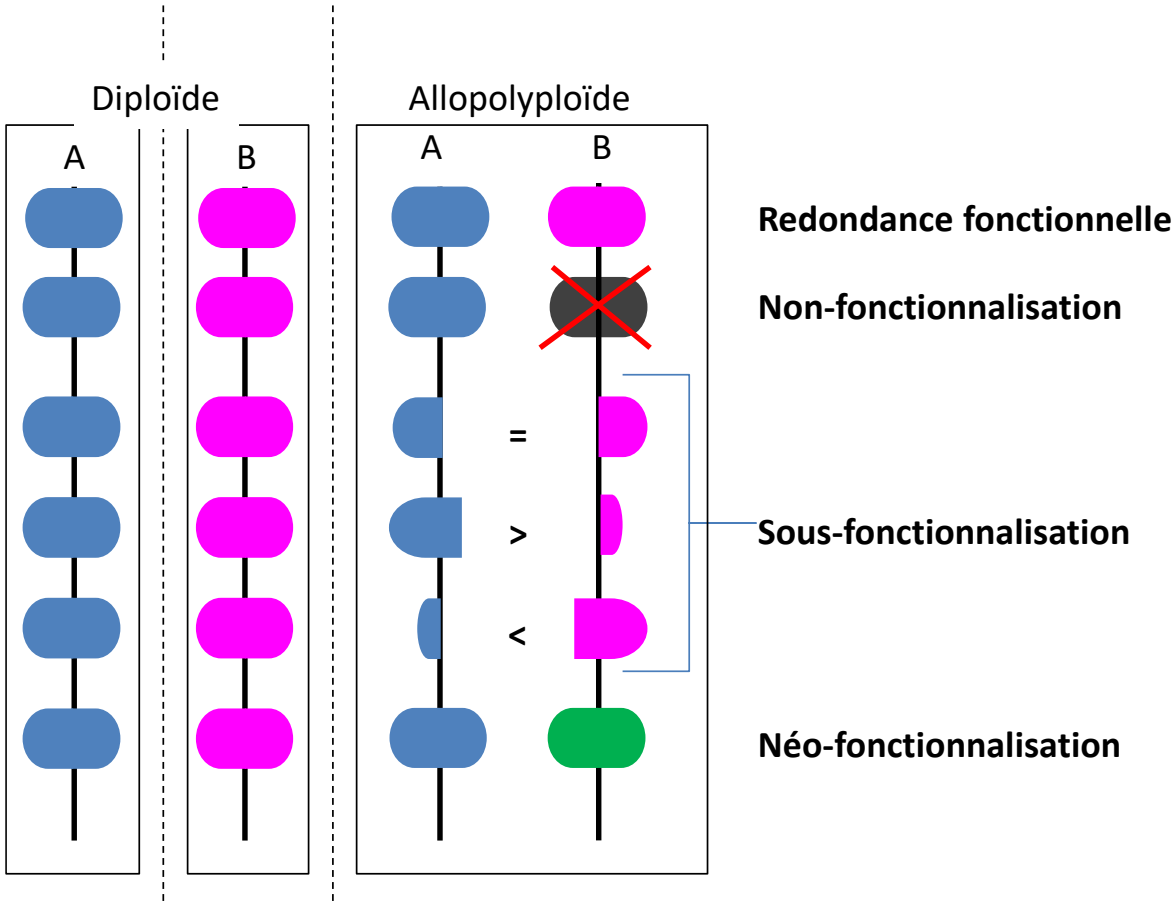


Figure 11: Devenir des gènes dupliqués.

Le maintient des deux copies dupliquées est une redondance fonctionnelle ; la non-fonctionnalisation consiste en la délétion d'une des copies ancestrales ; la sous-fonctionnalisation peut-être une contribution équivalente des copies ancestrales, ou biaisée vers l'un des deux parents ; la néo-fonctionnalisation est l'acquisition d'une nouvelle fonction.

Ces modifications épigénétiques des gènes régulateurs clefs chez les hybrides et les allopolyploïdes peuvent altérer les réseaux complexes de régulation de la physiologie et du métabolisme (Chen, 2013), induisant une reprogrammation de l'expression des gènes.

1.1.4.4. Le devenir des gènes dupliqués

La polyploïdie conduit à une duplication de la majorité des gènes du génome (Doyle et al., 2008). Plusieurs études ont porté sur les effets évolutionnaires de la polyploïdie et le devenir des gènes dupliqués (Muntzing, 1936); (Comai, 2005); (Otto, 2007); (Soltis and Soltis, 2009). Certains scientifiques considèrent la polyploïdie comme une impasse évolutionnaire (Stebbins, 1950 ; Arrigo and Barker, 2012), à l'inverse d'autres chercheurs considèrent la polyploïdie comme un impact majeur dans l'évolution et la diversification des espèces (Chen, 2010 ; Mayfield et al., 2011).

En 1970, Ohno propose le terme de non-fonctionnalisation, qui correspond à la délétion d'une copie de gène dupliquée (Fig. 11). Il suggère qu'une seule copie du gène suffit pour assurer la fonction du gène, la redondance apportée par la duplication entraîne le relâchement de la pression de sélection sur la deuxième copie, favorisant ainsi l'accumulation de mutations puis la perte de la fonction du gène. Cette copie devient un pseudogène ou peut aussi être entièrement délétée.

Bien qu'une quantité plus importante de produit des gènes puisse être avantageuse, il est suggéré que deux gènes ayant la même fonction (redondance fonctionnelle) ne sont pas stables au sein d'un génome sur le long terme de son évolution (Nowak et al., 1997). Par ailleurs, deux copies similaires ont plus de chance d'être maintenues si leurs fonctions diffèrent (Nowak et al., 1997). Si ces copies dupliquées évoluent vers un partage de la fonction, et participent conjointement à la fonction (chacune des copies se spécialisant dans une des sous-fonctions originelles), il s'agit dans ce cas d'une sous-fonctionnalisation (Lynch and Force, 2000 ; Cusack and Wolfe, 2007).

Par exemple, un gène ubiquitaire se duplique et chaque copie devient tissu spécifique en conservant la fonction originelle: c'est le modèle Duplication-Dégénération-Complémentation (DDC) proposé par Force (Force et al., 1999). La conservation par DDC des deux copies d'un gène s'établit lorsque chacune des copies a perdu l'une des fonctions du gène original. Chaque copie est indispensable à l'organisme.
La sous-fonctionnalisation a été révélée chez différents allopolyploïdes, par des biais d'expression des homéologues, chez le coton (Chaudhary et al., 2009), *Arabidopsis* (Blanc and Wolfe, 2004; Groszmann et al., 2011), *Tragopogon mirus* (Buggs et al., 2010), la paramécie *Paramecium tetraurelia* (Arnaiz et al., 2010), la *Xenopus laevis* (Semon and Wolfe, 2008), les humains (Gu et al., 2002), le blé (Zhang et al., 2011) et les espèces *Brassica* (Cheng et al., 2014).

Les gènes dupliqués peuvent aussi évoluer vers l'émergence d'une nouvelle fonction pour une copie, alors que l'autre copie maintient la fonction d'origine: une des deux copies d'un gène, en absence de pression de sélection, va pouvoir accumuler des mutations aléatoirement et acquérir ainsi une nouvelle fonction, c'est la néo-fonctionnalisation (Lynch and Force, 2000 ; Blanc and Wolfe, 2004 ; Chen, 2007 ; Lu et al., 2012).

La rétention ou la perte préférentielle de gènes dupliqués apparait dans un grand nombre de taxons[4] (Doyle et al., 2008) et reste similaire entre les polyploïdes récents et anciens, d'une même famille de plantes (Buggs et al., 2012). Des analyses d'EST et de séquences de génome montrent que le rapport rétention/perte de copies de gènes dupliqués n'est pas aléatoire. Des gènes sont dupliqués et re-dupliqués, alors que d'autres retournent itérativement à l'état singleton (Chapman et al., 2006), ceci concerne des gènes spécifiques de catégories fonctionnelles qui retiennent ou perdent préférentiellement des copies (Blanc and Wolfe, 2004). Les génomes d'angiospermes qui sont actuellement entièrement séquencés montrent des modèles non aléatoires de rétention et de perte de gènes suite à la duplication du génome (Adams and Wendel, 2005; Doyle et al., 2008 ; Brenchley et al., 2012 ; Chalhoub et al., 2014).

[4] Ensemble d'êtres vivants partageant certaines caractéristiques, à partir desquelles est établie leur classification.

Après la duplication du génome, les polyploïdes sont dans une phase « révolutionnaire » de changements génétiques et épigénétiques, suivie d'une phase « évolutionnaire » au cours de laquelle les polyploïdes acquièrent des modifications plus lentes (Levy and Feldman, 2002). Par accumulation de réarrangements et de délétions de séquences, les polyploides évoluent progressivement vers un état diploïde donnant lieu à de nouveaux diploïdes ou des paléopolyploïdes (Blanc and Wolfe, 2004; Levy and Feldman, 2004; Feldman and Levy, 2009 ; Hufton and Panopoulou, 2009; Tate et al., 2009; Buggs et al., 2010; Xiong et al., 2011).

Ce processus de diploïdisation serait un processus pour fixer des interactions inter-génomiques positives et maintenir l'asymétrie génomique résultant de la perte ou de la répression de gènes (Feldman et al., 2012).

Le maintient de gènes dupliqués n'est pas aléatoire et favorise un génome plutôt qu'un autre. Une telle asymétrie génomique se manifeste dans le blé allopolyploïde par le contrôle de différents caractères morphologiques et agronomiques, impliqués dans la production de rRNA, de protéines de réserve, voir dans l'interaction avec des agents pathogènes (Feldman et al., 2012).

Chez le blé, les espèces tétraploides *T. turgidum* et hexaploïdes *T. aestivum* portent le locus *Ph1* (Pairing homoeologous 1), qui contrôle la formation de bivalents exclusivement entre chromosomes homologues à la métaphase I de la méiose. Le locus *Ph1*, situé sur le bras long du chromosome 5B, est un des gènes inhibant l'appariement des chromosomes homéologues (Riley and Chapman, 1958), conférant aux blés polyploïdes un comportement diploïde. Le locus *Ph1* est une région composée d'un complexe de sept Cycline kinase dépendante (Cdk)-like gènes interrompus par de l'hétérochromatine sub-télomérique (Griffiths et al., 2006 ; Al-Kaff, 2008). Les régions homéologues sur les chromosomes 5A et 5D contiennent cinq et deux Cdk-like gènes respectivement.

L'organisation spécifique de ce locus sur le chromosome 5B a été observée chez tous les blés polyploïdes mais n'est présente dans aucune autre espèce diploïde des genres *Triticum* et *Aegilops*, alors qu'elle est présente chez *T. timopheveii* (2n=4x=28, AAGG), un autre blé allotétraploïde naturel. Le locus *Ph1* serait apparu peu après l'évènement de polyploïdisation et serait maintenu dans différents allopolyploïdes naturels de blé (Griffiths et al., 2006).

Figure 12: Origine du blé.

Le croissant fertile couvre le Liban, la Syrie, le sud de la Turquie, l'Iran, l'Irak, la Jordanie et Israël. (Source : Map courtesy of the university of Texas Libraries.)

1.2. Le blé: une culture d'importance économique et un modèle d'étude de la polyploïdie

1.2.1. Une espèce d'intérêt socio-économique majeur

Le blé, une des premières espèces cultivées puis domestiquée, est originaire du croissant fertile du Proche-Orient (actuels Liban, Syrie, Sud de la Turquie) (Fig. 12). Sa domestication[5] a permis l'émergence des premières civilisations (babyloniennes, assyriennes, égyptiennes) jusqu'aux civilisations occidentales, due à sa facilité de culture à grande échelle et à son stockage pour de longues durées.

Avec le riz et le maïs, le blé fournit plus de la moitié des calories et protéines absorbées quotidiennement. Avec plus de 215 millions d'hectares (Ha) semés chaque année, le blé est la céréale la plus cultivée dans le monde (17% des terres cultivées environ). La production du blé en 2013 atteint près 21 000 kilos par seconde, soit 653 millions de tonnes par an (2013) (dont 95 % correspondent à du blé tendre et 5 % à du blé dur). Actuellement, la France est au premier rang des 27 pays de l'Union européenne en production de céréales (blé tendre, blé dur, orge, maïs, avoine, seigle, sorgho, triticale) et sur le plan mondial, elle est au 5ème rang après la Chine, l'Inde, la Russie et les Etats-Unis (USDA, 2012).

Le blé est aussi au premier rang dans les échanges agroalimentaires internationaux. Dans cette production et ces transactions, la France détient un rôle important, car en tant que premier producteur de blé européen elle exporte près de 50% de sa récolte dans une centaine de pays. Le blé, utilisé pour l'alimentation (farine, céréales du petit déjeuner, pâtes, fermentation de la bière), matériau de construction (toitures de chaume) ou encore pour l'alimentation du bétail, représente, en France, 54 milliards d'euros de chiffre d'affaires. La production mondiale, en progression constante, et les échanges qui se multiplient entre les différentes régions du monde font de cette céréale l'une des principales valeurs de l'économie mondiale.

[5] La domestication d'une espèce animale ou végétale est l'acquisition et la transformation de caractères et de comportements héréditaires au contact de l'homme, que ce soit suite à une interaction prolongée ou à un effort volontaire de sélection.

Le besoin d'augmentation de la production agricole, d'ici à 2050, nécessiterait une croissance annuelle d'environ 2 % (Dixon et al., 2009 ; Feuillet et al., 2011). L'impact de l'agriculture sur l'environnement doit-être réduit, et des variétés plus tolérantes à la sécheresse, plus résistantes aux maladies fongiques, aux virus et aux insectes doivent être sélectionnées (Feuillet et al., 2011). Pour y répondre, il faut développer de nouvelles méthodes stratégiques, de nouveaux outils et ressources (telle que la séquence du génome) pour une meilleure efficacité et une plus grande précision dans la sélection de variétés (Feuillet et al., 2011).

L'intégralité des informations, dont les séquences du génome, permettront entre autre de développer de très nombreux marqueurs moléculaires, pour accélérer les programmes de sélection variétale et l'identification de gènes impliqués dans les caractères d'intérêt agronomique (Feuillet et al., 2011).

1.2.2. Les enjeux du séquençage du génome du blé

Le génome du blé à pain, d'une taille de 17 000Mb distribuées sur 21 chromosomes, est 5 fois plus grand que le génome du maïs et 40 fois plus grand que celui du riz. La complexité de son génome hexaploïde (composé de 3 sous-génomes homéologues A, B et D) très riche en éléments transposables (85%) rend son analyse moléculaire et son séquençage difficiles et coûteux.

Plusieurs initiatives de séquençage du génome du blé ont été lancées (Brenchley et al., 2012 ; Jia et al., 2013 ; Luo et al., 2013). La création d'un consortium international pour le séquençage du génome de blé (International Wheat Genome Sequencing Consortium, http://www.wheatworld.org/research/major-wheat-research-projects/international-wheat-genome-sequencing-consortium/) a permis de débuter le projet en 2005 et a abouti récemment à des premiers résultats publiés (Choulet et al., 2014 ; IWGSC, 2014 ; Marcussen et al., 2014 ; Pfeifer et al., 2014). Ces projets permettent un accès public aux séquences du génome, plus au moins bien analysées, annotées et alignées sur des cartes génétiques, ce qui va accélérer la création et/ou le déroulement de projets innovants. Ainsi, il sera possible d'isoler les gènes d'intérêt pour identifier leurs fonctions biologiques et accroître les rendements de production. Le séquençage du génome permettra de favoriser la sélection de nouvelles variétés de blé, et leurs améliorations.

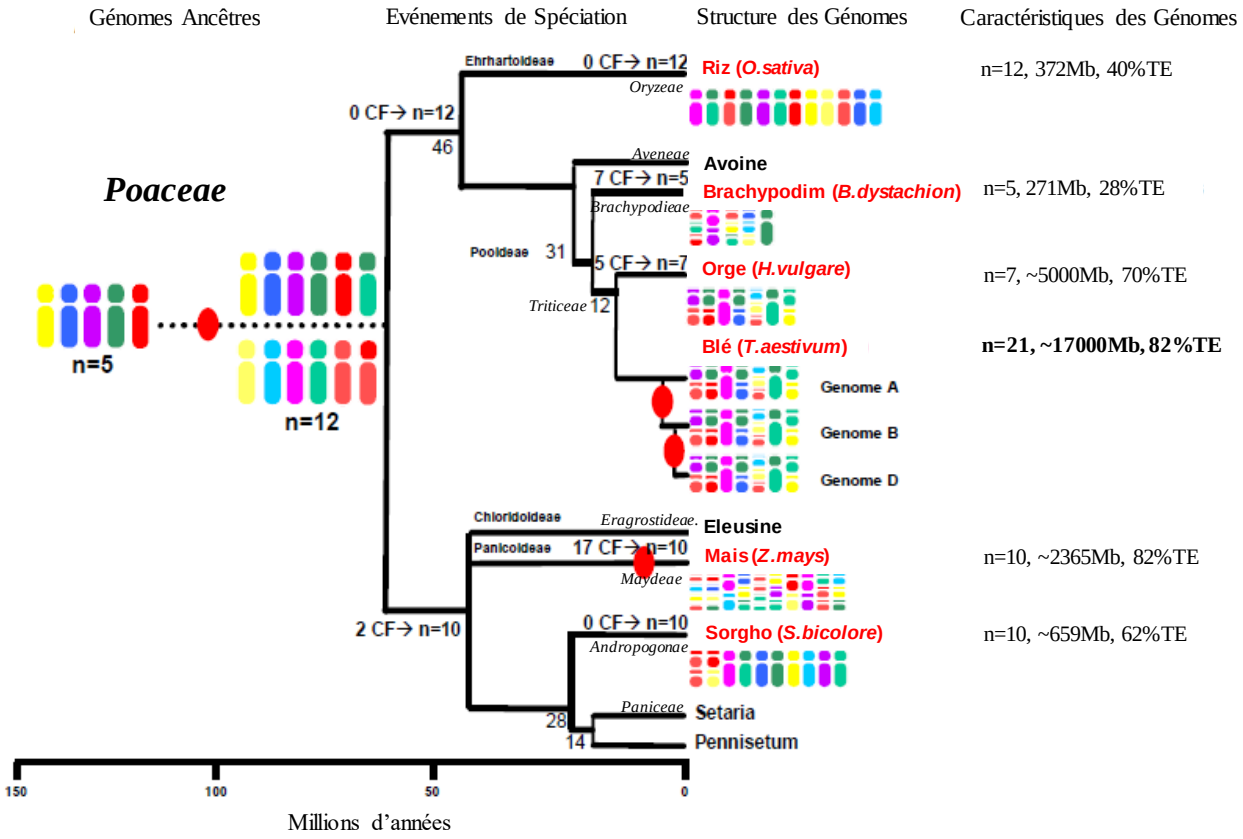


Figure 13: Modèle d'évolution des espèces de la famille des Poaceae et de la conservation des gènes.

Figure inspirée de (Pont et al., 2011). Les proportions en TE des différents génomes sont d'après (Paterson and al., 2009) pour le sorgho, le mais, le riz ; d'après (Vogel, 2010) pour *B.distachyon*, (Paux et al., 2006) et (Charles et al., 2008) pour le blé ; et (Wicker et al., 2009) pour l'orge. Les événements de polyploïdisation sont indiqués par des cercles rouges. L'évolution du nombre de chromosome des espèces à partir de la structure du génome ancêtre est indiquée avec le nombre d'évenements de fusion de chromosome (CF).

Afin de réduire la complexité d'analyse, ces travaux de séquençage seront affinés sur chaque bras de chromosome individuellement.

Différemment des OGM (organismes génétiquement modifiés), ou quelques gènes ont été introduits ou modifiés, les chercheurs visent préférentiellement les réseaux de gènes, qui concernent l'intégralité d'une voie métabolique, dans des blés anciens ou sauvages proches des blés domestiques pour l'amélioration et l'introduction de nombreux caractères utiles comme la résistance à des maladies, à la sécheresse, etc… (Huet, 2013).

1.2.3. Evolution et organisation des génomes du blé

1.2.3.1. Taxonomie du blé parmi les *Poaceae*

Les différentes espèces de blé appartiennent aux deux genres *Triticum* et *Aegilops* qui sont, selon la classification hiérarchique des espèces; des angiospermes (plantes à fleurs) monocotylédones de la famille des *Poaceae* (poacées en français, anciennement graminées), de la sous-famille des *Pooideae* et de la tribu des *Triticeae* (Fig. 13).

La famille des *Poaceae* compte plus de 600 genres et 10 000 espèces, poussant sous des latitudes et des climats diversifiés (Kellogg, 2001). De nombreuses espèces de cette famille ont été domestiquées et représentent un intérêt agronomique majeur : le riz (genre *Oryza*), le maïs (genre *Zea*), le sorgho (genre *Sorghum*), l'avoine (genre *Avena*), le seigle (genre *Secale*), l'orge (genre *Hordeum*) et le blé (genres *Triticum* et *Aegilops*).

Différentes classifications des *Triticeae* ont été proposées et varient en fonction des critères (morphologiques, cytologiques et moléculaires) utilisés pour différencier ces espèces. Depuis 1998, la classification de référence est celle réalisée par « Angiosperm Phylogeny Group » (APG). La classification pylogénétique APG III dépend principalement de trois marqueurs moléculaires (deux chloroplastiques et un mitochondriale) (APG 1998 ; APG II, 2003 ; APG III, 2009 ; Chase and Reveal, 2009 ; Haston, 2009).

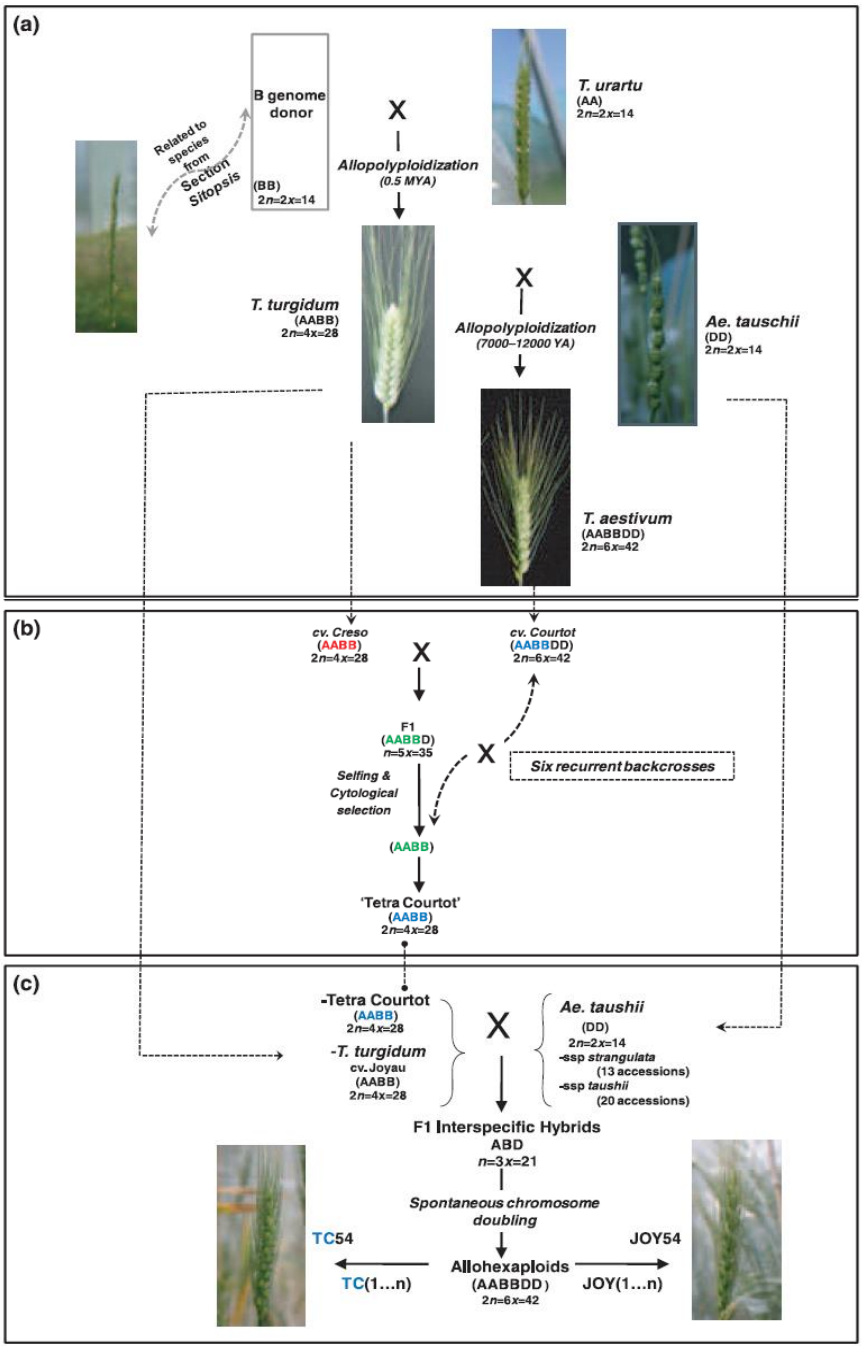


Figure 14: Evolution et origine de blé allohexaploide naturel et synthétique.

(a) Évolution des espèces de blés (genre *Triticum* et *Aegilops*) à travers les évenements d'allopolyploïdisation, d'après (Kihara, 1944; McFadden and Sears, 1946 ; Feldman et al., 1995 ; Nesbitt and Samuel, 1996 ; Blake et al., 1999 ; Huang et al., 2002a). (b) Représentation schématique de l'extraction du composant 'Tetra Courtot' (2n=4x=28, BBAA) du blé à pain Français (allohexaploïde) cv Courtot (Courtot ⁄ cv Creso ⁄ ⁄ 6* Courtot), selon Kerber (1964). (c) Les blés allohexaploïdes nouvellement synthétisés ont été obtenus par hybridation entre deux génomes progéniteurs AB et 33 accessions d'*Aegilops tauschii* (2n=2x=14, DD), suivi d'un doublement spontané des chromosomes, d'après (Mestiri et al., 2010).

1.2.3.2. Evolution des génomes du blé parmi les *Poaceae*

L'étude des duplications dans les génomes confirment l'existence d'un événement de paléotétraploïdisation (duplication complète du génome), commun à toutes les *Poaceae*, il y a 50-70 Ma (Tang et al., 2008). Suite à cet événement de polyploïdisation ancestral, qui semble avoir concerné un génome possédant un ensemble de 5 chromosomes ancestraux, les *Poaceae* ont divergé grâce à des évènements de translocations et de fusion de chromosomes pour atteindre un ancêtre commun (n=12 chromosomes) aux monocotylédones modernes (Salse, 2008a ; Devos, 2010 ; Pont et al., 2011 ; Murat et al., 2014) (Fig. 13). Un événement de duplication totale du génome, plus ancien, aurait eu lieu chez les monocotylédones peu après leur divergence avec les dicotylédones (estimé à plus de 200 Ma, mais reste à préciser) (Tang, 2008b).

En plus de ces événements de polyploïdisation anciens, communs aux Poaceae (Salse, 2008a ; Pont et al., 2011 ; Murat et al., 2014), les différentes espèces du blé appartenant aux genres *Triticum* et *Aegilops* ont subi, au cours de leur évolution, différents événements de polyploïdisation relativement récents aboutissant à l'apparition de nombreuses espèces tétraploïdes et hexaploïdes dont les blés cultivés de nos jours, le blé dur (*T. turgidum ssp. durum*, 2n=4x=28, BBAA) et le blé tendre (*T. aestivum* ssp. *aestivum*, 2n=6x=42, BBAADD).

L'amidonier sauvage dur est une espèce tétraploïde *Triticum turgidum ssp. dicoccoides* (2n=4x=28, BBAA) (Dvorak and Zhang, 1990 ; Dvorak et al., 1993) (Fig. 14).

Selon le premier modèle de l'histoire phylogénétique des blés, le blé tétraploïde provient d'un évènement de polyploïdisation, il y a environ 500 000 ans (Dvorak, 2005 ; Chalupska et al., 2008), entre deux espèces diploïdes *Triticum urartu* (2n=2x=14, AA) donneur du génome A et une espèce encore non identifiée de la section *Sitopsis*, donneur du génome B (Feldman et al., 1995 ; Blake et al., 1999 ; Huang et al., 2002b ; Dvorak, 2006). Le donneur du génome A est bien caractérisé comme étant l'espèce sauvage diploïde *T.urartu*, mais le donneur du génome B reste à identifier. La section *sitopsis* des *Aegilops* (génome SS) regroupe cinq espèces, des travaux de phylogénie et de cytogénétique indiquent que *Ae. speltoides* (2n=2x=14, SS) serait l'espèce la plus proche du donneur du génome B (à 60%) (Vedel, 1981 ; Raskina, 2002 ; Salse et al., 2008b). Le blé tendre cultivé est une espèce hexaploïde *Triticum aestivum* (2n=6x=42, BBAADD). Cette espèce provient de l'hybridation d'une forme domestiquée du blé tétraploïde *T. turgidum ssp dicoccum* appelée amidonnier domestique, avec le diploïde *Ae. tauschii* donneur du génome D.

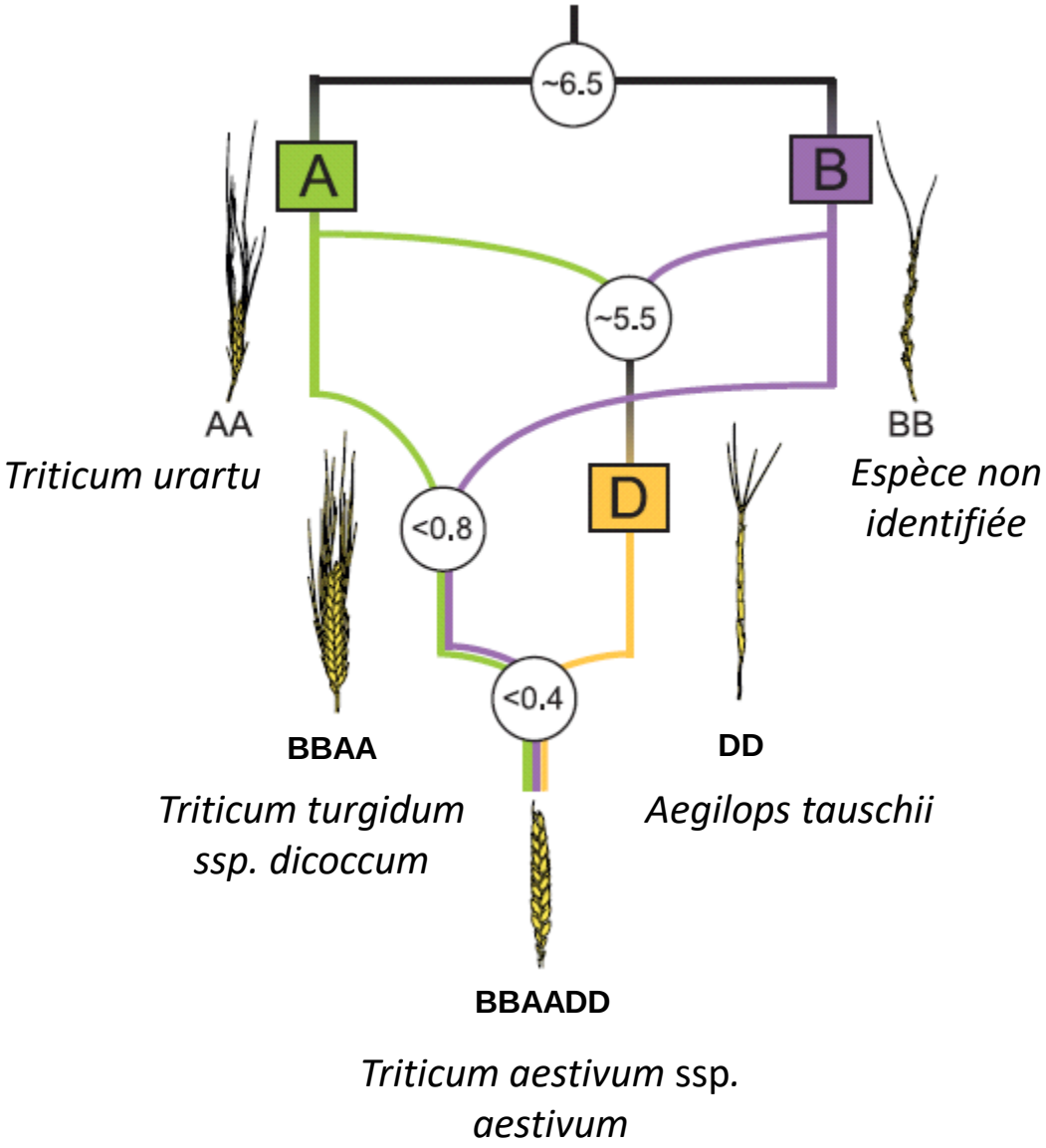


Figure 15: Modèle de l'histoire phylogénétique du blé à pain (*Triticum aestivum*, BBAADD).

D'après (Marcussen et al., 2014).

Des travaux très récents proposent un nouveau modèle de l'histoire phylogénétique du blé à pain (Marcussen et al., 2014). Selon ce nouveau modèle, les génomes A et B divergent d'un ancêtre commun il y a environ 6.5 millions d'années. Une première hybridation aurait eu lieu, il y a environ 5.5 millions d'années, entre les génomes A et B et aurait conduit à l'espèce *Ae. tauschii* porteur du génome D par une spéciation hybride homoploïde[6]. Puis, un évenément de polyploïdisation entre le génome (AA) et le génome (BB) a eu lieu il y a 0,8 millions d'années et a conduit au blé dur tétraploïde. Le blé à pain serait issu d'une deuxième allopolyploïdisation, entre le blé dur et *Aegilops tauschii* (DD), il y a 0,4 millions d'années (Fig. 15). Des études portant sur la conservation des gènes le long des chromosomes homéologues des génomes du blé viennent confirmer ce modèle, en affirmant que le contenu en gènes des chromosomes A et B est plus semblable à celui des chromosomes du génome D, plutôt qu'entre eux (IWGSC, 2014).

Les ressources génomiques, génétiques et cytogénétiques développées pour l'étude du blé et de ses génomes parentaux, en font un excellent système pour étudier l'allopolyploïdie. La possibilité de « synthétiser » de nouveaux blés (*Triticum*) par croisement entre différents progéniteurs permet de caractériser, dès la première génération, l'effet direct de la polyploïdie sur l'évolution et la régulation de l'expression des gènes.

1.2.3.3. Importance des éléments transposables dans les génomes du blé

Les éléments transposables (TEs) sont des séquences d'ADN capables de se déplacer et de se multiplier de manière autonome dans un génome. Les TEs constituent l'essentiel de l'ADN répété (SanMiguel et al., 1998), et sont communs à tous les organismes vivants. Leur importance relative dans le génome varie grandement selon l'espèce.

La diversité et l'évolution des génomes du blé (du groupe *Triticum-Aegilops*) est déterminée, en partie, par l'activité des TEs qui constituent une forte proportion du génome (plus de 80 %) (Paux et al., 2006 ; Charles et al., 2008). Ce sont des éléments très importants et dynamiques des génomes du blé. Ils sont en partie responsables des variations de taille très importantes entre les différentes espèces de blé pouvant atteindre plusieurs centaines de Mb pour des espèces ayant le même niveau de ploïdie (Bennett and Smith, 1976 ; Bennett and Smith, 1991).

[6] La spéciation hybride homoploïde crée une nouvelle espèce tout en gardant un même niveau de ploïdie identique à celui des génomes parentaux.

Une description détaillée des TEs et de leur importance dans le blé a été présentée dans une thèse soutenue au laboratoire cette même année (Belcram, 2014). Je mets cette description que j'ai actualisée en Annexe 1.

La proportion des TEs varient selon les classes I et II. Les éléments de classe I représentent plus de 60% des génomes du blé, plus précisément 66% du chromosome 3B du blé (Choulet et al., 2014) comparativement au 10% du génome de la drosophile. Le séquençage du génome du blé révèle une différence dans la distribution des TEs dans les sous-génomes A, B et D (IWGSC, 2014): les TEs de classe I (les rétroéléments) sont plus abondants dans le sous-génome A comparé aux sous-génomes B et D (A>B>D). Les éléments de classe II sont près de 10 fois plus importants dans le génome du blé (20%) que celui du maïs (3.2%) (Schnable et al., 2009). Les travaux récents du séquençage du chromosome 3B montrent que 60% de la séquence est composée de TEs de classe II, qui sont majoritairement des éléments de la super-famille des CACTA (Choulet et al., 2014).

La prolifération des TEs n'est pas constante au cours de l'évolution des espèces, ni homogène le long des chromosomes. On distingue des périodes d'activité très fortes et des périodes d'activité plus faible. La plupart des TEs, qui forment le chromosome 3B se sont insérés avant la polyploïdisation (il y a environ 0.5 millions d'années (Ma)), et sont devenus moins actif par la suite (Choulet et al., 2014). Les événements récents d'insertions de TEs sont homogènes dans les régions distales (R1 et R3) et proximale (R2) du chromosome 3B (Annexe Fig. 3). Les insertions plus anciennes (>1.5 Ma) ont été plus rapidement éliminées dans les régions distales du chromosome 3B (Choulet et al., 2014). La densité de distribution des TEs le long du chromosome 3B n'est pas aléatoire, il y a une plus faible densité dans les régions distales R1 (73%) et R3 (68%), et une plus forte densité dans la région proximale R2 (88%). La région centromérique-péricentromérique montre une plus forte densité de TEs (93%) (Choulet et al., 2014). A l'échelle du génome entier, sous l'hypothèse d'une dynamique d'insertion/élimination similaire entre les 3 sous-génomes A, B et D, la distribution en rétrotransposons est plus faible dans le sous-génome B (IWGSC, 2014).

1.3. Les outils d’analyse du transcriptome

1.3.1. Évolution des techniques de mesure de l’expression des gènes

La quantification de l’expression des gènes intègre différents niveaux de régulation.
Le génome décrit l'ensemble du matériel génétique d'un individu ou d'une espèce codé dans son ADN. Il contient en particulier toutes les séquences codantes (transcrites en ARN messagers, et traduites en protéines) et non codantes (non transcrites, ou transcrites en ARN, mais non traduites).
Le transcriptome rassemble l'ensemble des ARN (messagers, ribosomiques, de transfert, non codant et autres types d'ARN) issus de la transcription du génome, exprimés donc dans une cellule à un instant donné et/ou dans une condition donnée. Il reflète l’état fonctionnel du génome.

Au cours de mes travaux de thèse, je me suis intéressée plus particulièrement à étudier la reprogrammation de l’expression des gènes induite par la polyploïdie en caractérisant le transcriptome des blés polyploïdes de façon comparative avec les espèces progénitrices.

La caractérisation et la quantification du transcriptome dans un tissu donné et dans des conditions données permettent d'identifier les gènes exprimés, de déterminer les mécanismes de régulation d'expression des gènes et d’évoluer les réseaux d'expression des gènes.

Les techniques de quantification de l’expression des gènes ont grandement évolué,: du Northern blot aux méthodes de génomiques telles que cDNA-AFLP jusqu’au séquençage des ARNs par les outils de nouvelles générations de séquençage (NGS) (Lee and Chen, 2001 ; Madlung et al., 2002 ; Hegarty et al., 2006 ; Tate et al., 2006 ; Wang et al., 2006b ; Flagel et al., 2008 ; Gaeta et al., 2009 ; Chague et al., 2010 ; Flagel and Wendel, 2010 ; Koh et al., 2010 ; Chelaifa et al., 2013 ; Yoo et al., 2013).

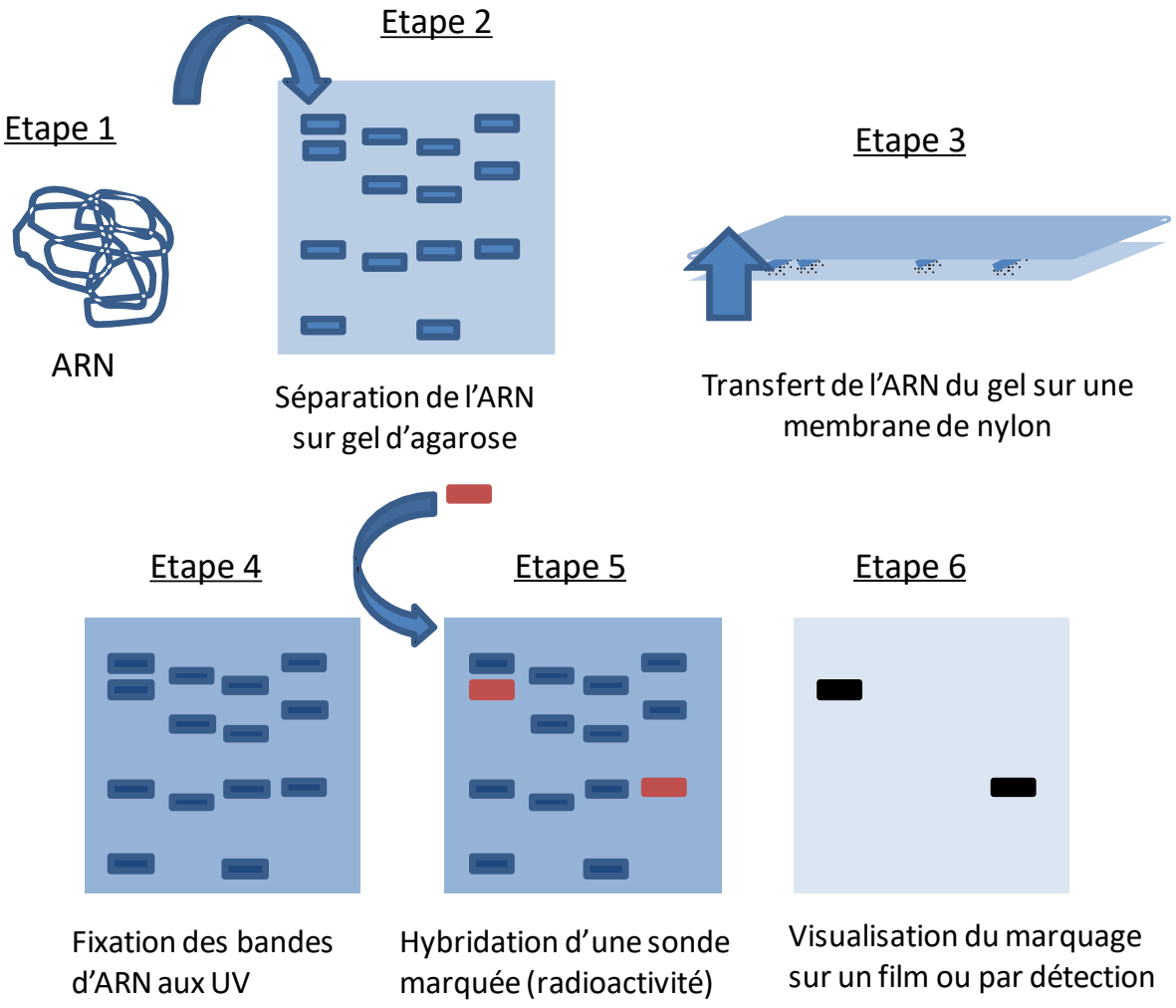


Figure 16: Méthode du Northern Blot.

(1) Extraction des ARNs, (2) séparation des ARNs sur gel en fonction de leurs tailles, on obtient par champ de migration un profil spécifique pour chaque condition expérimentale, (3) les profils sont transféré sur une membrane pour le marquage, (4 et 5) hybridation des sondes marquées sur les ARNs fixés sur la membrane, (6) révélation des ARNs étudiés et communs entre les différentes conditions expérimentales.
Inspiré de http://virology-microbiology-b.blogspot.fr/

Dans les paragraphes qui suivent, je décris et retrace les premières techniques de biologie moléculaire utilisées pour caractériser l'expression des gènes, puis je détaille les tecnhiques les plus récentes et plus couramment utilisées: les techniques à haut débit.

1.3.1.1. Les premières Techniques de biologie moléculaire

1.3.1.1.1. Northern blot

Le Northern blot est une des premières méthodes de biologie moléculaire permettant l'analyse de l'ARN (Alwine et al., 1977). Cette technique a été appelé Northern, par jeux de mots, pour signifier l'hybridation d'ARN par opposition à la technique similaire d'hybridation d'ADN appelé Southern blot (Southern, 1975) de laquelle elle est dérivée. Les fragments d'ARN sont analysés par électrophorèse, afin de les séparer selon leur taille, puis transférés par capillarité sur une membrane de nylon qui est hybridée avec une sonde spécifique marquée radioactive (sonde ADN ou ARN) du gène d'intérêt (Fig. 16).

À la différence du Southern Blot, qui utilise un traitement d'hydroxyde de sodium dans le gel d'électrophorèse comme dénaturant et qui dégraderait l'ARN, le traitement par formaldéhyde évite les repliements en structures secondaires des ARN. Cette technique permet d'évaluer la distribution des ARNs dans les tissus, d'observer les intermédiaires de maturation d'ARNm (ajout de la coiffe en 5' et de la queue polyA, épissage) et les différentes formes d'épissage des ARN.

Toutefois, cette méthode utilise des produits nuisibles tels que le formaldéhyde (hautement toxique), des marquages radioactifs, le bromure d'éthidium (mutagène), le DEPC (diethylpyrocarbonate: inhibiteur RNase) et les lampes UV.

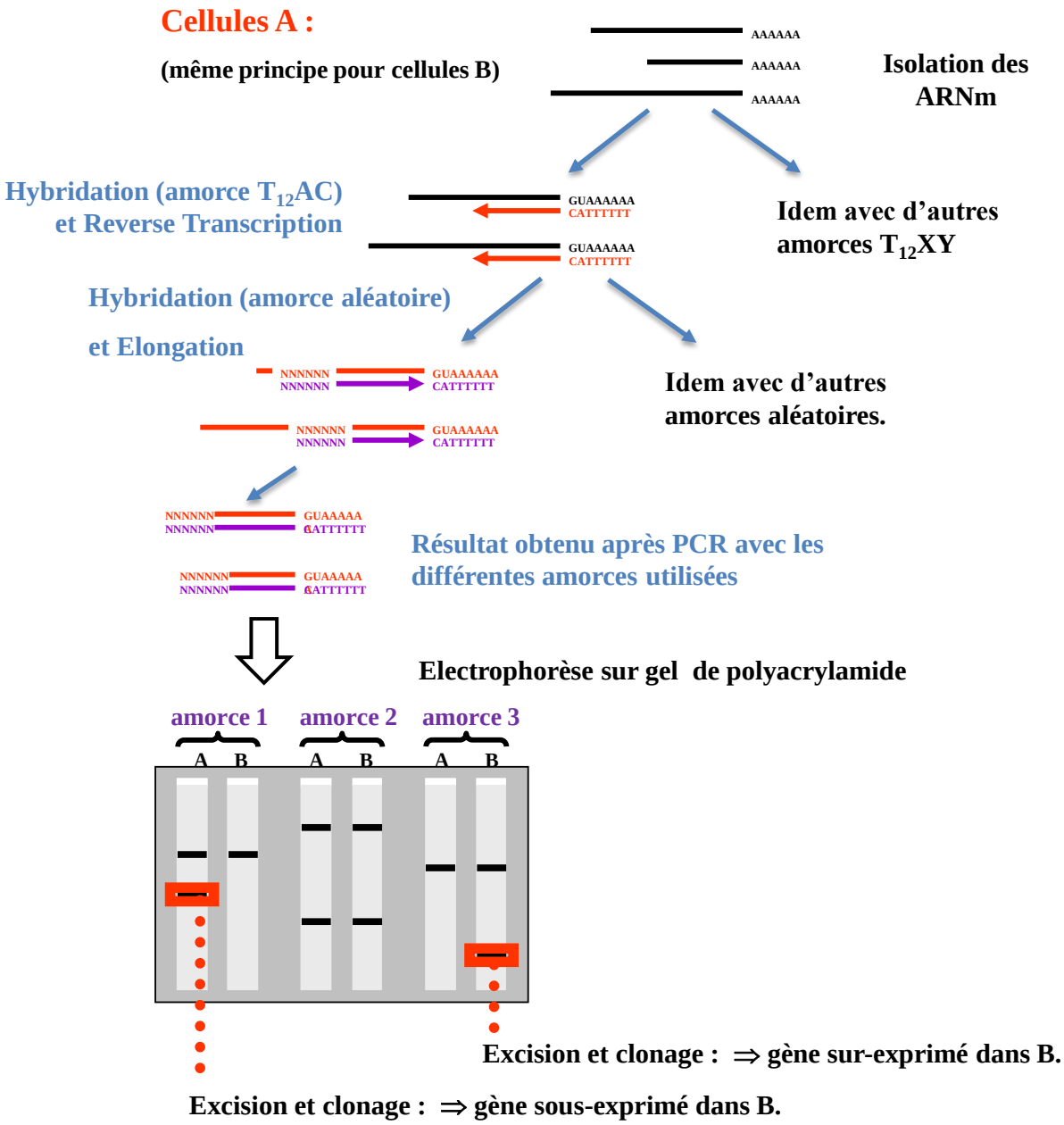


Figure 17: Principe de la DD-RT-PCR.

1.3.1.1.2. DD-RT-PCR

La DD-RT-PCR (Differential Display Reverse Transcription-PCR) est une technique de tri et de comparaison des profils d'ARNm (Liang and Pardee, 1992) qui permet de repérer des différences d'expression de gènes entre deux conditions (Fig.17). Ces gènes sont ensuite isolés et séquencés si nécessaire. Cette technique est fondée sur trois étapes majeures :

(i) la transcription inverse à partir d'ARNm aboutissant à l'obtention d'ADNc monocaténaires,

(ii) la synthèse du brin complémentaire et l'amplification par PCR des ADNc,

(iii) la séparation éléctrophorétique des fragments d'ADNc bicaténaires avant leur clonage.

Ces trois étapes sont réalisées simultanément et en parallèle sur des ARNm isolés de différents tissus ou traitements. L'analyse des électrophorèses sur gel permet de comparer et d'identifier des bandes d'ADNc différentiellement accumulées suivant les différents tissus ou traitements (Tagu and Moussard, 2003).

Après une extraction des ARN totaux, incluant les ARNm polyadénylés, la transcription inverse permet de synthétiser les ADNc monocaténaires, à l'aide d'amorces $T_{12}XY$, complémentaires à la queue poly(A) (X est une quelconque base sauf une thymidine, Y correspond à toute base). Il existe 12 combinaisons possibles d'amorces, toutes complémentaires à des familles d'ARNm différents. Ce choix d'amorce détermine le tri des ARNm analysés, et permet une meilleure résolution du gel d'éléctrophorèse qui suit. En moyenne, ce tri permet de séléctionner un douzième de la population d'ARN poly(A) (Tagu and Moussard, 2003).

Les ADNc monocaténaires obtenus sont amplifiés par PCR, avec deux amorces : une amorce $T_{12}XY$ complémentaire à l'extrémité 3' de l'ADNc (identique à celle utilisée pour la transcription inverse), et une amorce complémentaire à l'extrémité 5' qui serait une amorce aléatoire de 10 paires de bases. Cette amplification aboutit à des fragments d'ADNc bicaténaires de tailles différentes (Tagu and Moussard, 2003).

Ces ADNc bicaténaires sont séparés sur gels d'acrylamide, dont la résolution se fait à la base près pour des fragments de 500pb. Si des ARNm sont différentiellement accumulés entre deux conditions, il est possible d'exciser les bandes différentiellement exprimés.

Une trop faible quantité d'ADNc, dans une bande, ne peut être directement sous-clonée, une réamplification de cet ADNc avec le couple d'amorces précédent est nécessaire. Ce fragment d'ADNc est cloné via un plasmide et séquencé. A partir de cette séquence d'ADN, une sonde est générée afin de mesurer l'expression du gène correspondant, par criblage d'une banque d'ADNc ou une banque d'ADN génomique (Tagu and Moussard, 2003).

Toutefois, cette technique présente des limites, on observe notamment une saturation pour les gènes fortement transcrits et donc une sous-estimation de la bande, ce qui engendre beaucoup de faux positifs dans les fragments de PCR clonés. Aussi, une bande peut correspondre à plusieurs fragments d'ADNc différents de même longueur.

1.3.1.1.3. AFLP-ADNc

La technique AFLP-ADNc (Amplified Fragment-Length Polymorphism ou Polymorphisme de Longueur des Fragments Amplifiés) (Bachem, 1996) permet de séparer les fragments d'ADNc en sous-groupes. La technique est similaire à la DD-RT-PCR de part la synthèse des ADNc et la comparaison des profils entre conditions afin de détecter des gènes différentiellement exprimés. Les ADNc sont digérés par deux enzymes de restriction différentes, puis différentes combinaisons d'adaptateurs sont liguées aux extrémités des sites de coupure des fragments. Ces fragments sont amplifiés par PCR avec des couples d'amorces spécifiques des différentes combinaisons. La spécificité des différentes combinaisons d'amorces permet d'obtenir des sous-groupes d'amplifias, qui sont séparés par électrophorèse sur gel. Les profils obtenus sont comparés entre sous-groupes de même combinaison d'amorces de conditions expérimentales différentes (Farce, 2000).

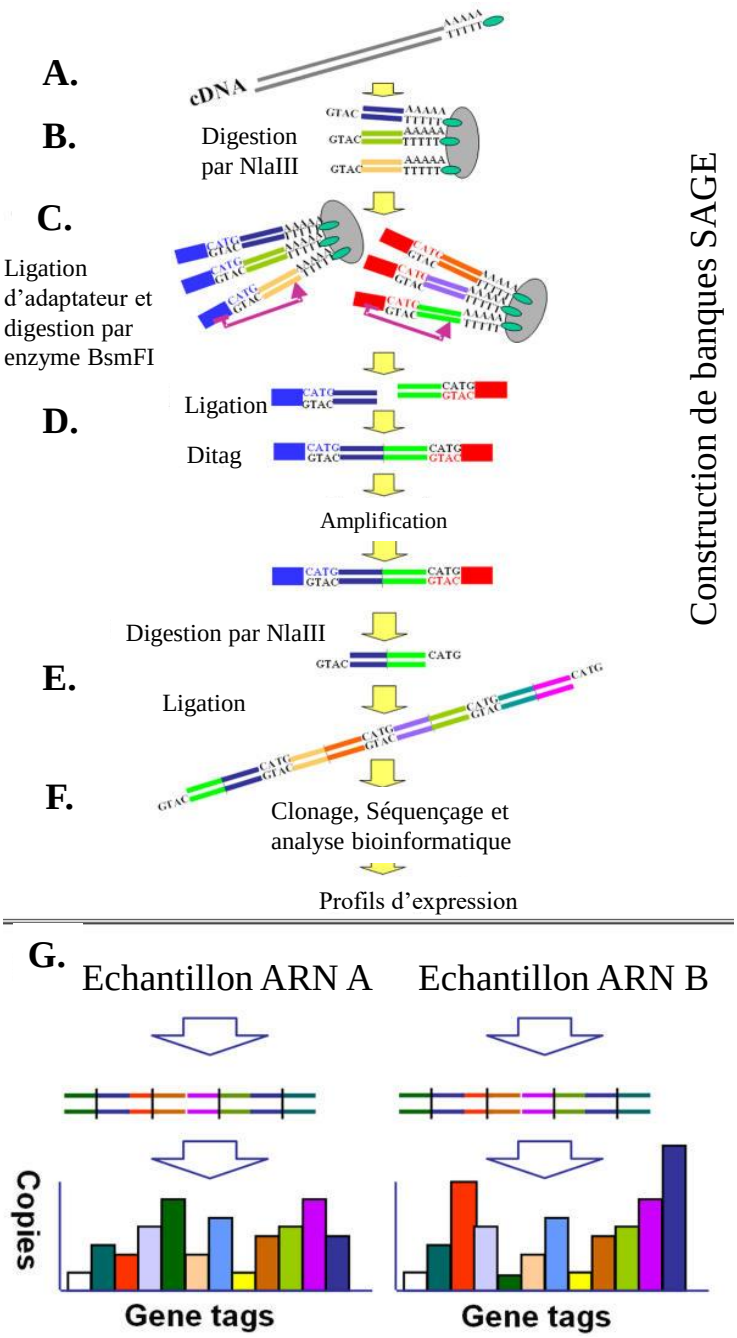


Figure 18: Principe de la technique SAGE.

(A) une population d'ARN reverse transcrits en ADNc utilisant des amorces attachées à des billes magnétiques. (B) les cDNAs sont digérés par l'enzyme de restriction Nla III. (C) Des adapteurs contenant la séquence reconnue par BsmF I sont liés pour découper le cDNA. Les etiquettes sont libérés des billes par BsmFI (qui coupe à quelques bases en aval de son site de reconnaissance). (D) les etiquettes d'ADN libérées sont liées ensemble pour former des "ditags". (E) Les ditags sont amplifiés puis digérés par Nla III pour retirer les adaptateurs. (F) Les ditags sont liés tous ensemble pour former une seule séquence qui est ensuite clonée dans un vecteur plasmidique pour générer la banque SAGE. (G) les différents niveaux d'expression des gènes sont déduits par les proportions des séquences tags comptés, d'après (Garnis et al., 2004).

1.3.1.1.4. SAGE et ses dérivés

SAGE (Serial Analysis of Gene Expression) est une technique d'analyse quantitative et qualitative en série de l'expression de gènes (Fig. 18). L'ensemble des ARN est converti en ADNc, puis un traitement enzymatique permet d'isoler à partir des extrémités une courte séquence spécifique de 9 à 14pb dit « tag » permettant d'identifier le gène dont elle dérive. Une cinquantaine d'étiquettes ou tag sont concaténées en un seul fragment d'ADN, puis séquencées pour identifier chaque étiquette. Les gènes, à l'origine des tags, peuvent alors être identifiés par comparaison de la séquence avec les banques. À partir de ces séquences, il est possible d'évaluer la fréquence des différents transcrits dans un échantillon donné. Cette technique n'est pas adaptée pour l'étude de nombreux échantillons, et nécessite l'existence de bases de données de séquences génomiques complètes pour les espèces étudiées.

Plusieus techniques dérivent de la méthode SAGE (Fig.19) :

- longSAGE: la méthode longSAGE (Saha, 2002) emploie une enzyme de restriction, qui coupe entre 18 et 20 pb en aval du site de reconnaissance. Les fragments découpés deviennent des étiquettes de 18-20 pb plus spécifiquement alignable avec les génomes de référence. Les étiquettes issues de la méthode longSAGE obtenues sont spécifiques en 5' ou en 3' des fragments d'ADNc (méthodes : 5′ longSAGE ou 3′ longSAGE).

- CAGE : la méthode CAGE (Kodzius et al., 2006) (Cap Analysis of Gene Expression) identifie, précisément, les sites de début de transcription en 5' des gènes et produit des étiquettes d'une longueur de 20-21 pb. Cette méthode permet d'étudier la structure des régions promotrices, et donc une analyse de l'ensemble des promoteurs du génome.

- SuperSAGE : Cette méthode (Matsumura and Kruger, 2005) est basée sur la méthode SAGE, elle génère des étiquettes plus longues de 26 paires de base. Chaque base supplémentaire allongeant la taille de l'étiquette augmente la précision de l'annotation des transcrits. Cet accroissement de la taille diminue la possibilité qu'une étiquette soit présente sur deux ARN différent.

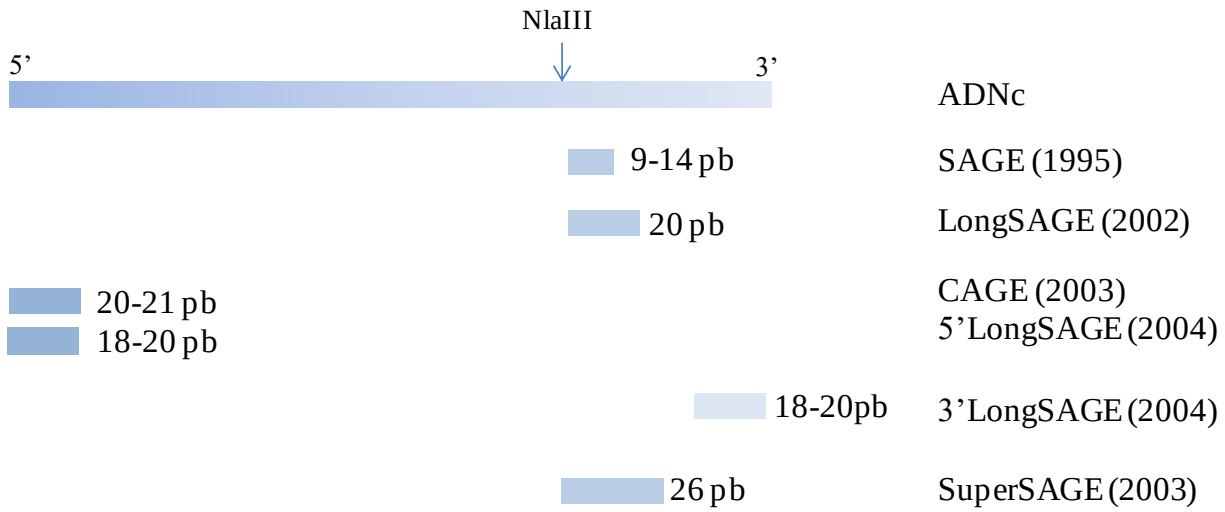


Figure 19: La méthode SAGE et ses dérivés.

L’étiquette SAGE extrait une séquence, en 3’, entre 9-14pb en aval du site de reconnaissance de NlaIII. LongSAGE utilise l’enzyme MmeI pour générer des etiquettes de 20pb. Les étiquettes CAGE et 5′ LongSAGE sont dérivées de l’extrémité 5′ des fragments d’ADNc, les étiquettes 3′ LongSAGE sont dérivées de l’extrémité 3’. SuperSAGE génère des étiquettes de 26pb, d’après (Fullwood et al., 2009).

1.3.1.2. Les Techniques à haut débit

1.3.1.2.1. Puces à ADN

Les puces à ADN (ou biopuces), mises au point au début des années 1990, regroupent plusieurs disciplines: l'électronique, la chimie, l'analyse d'images et l'informatique.

Cette technique est fondée sur le principe d'hybridation selon lequel deux fragments d'ADN complémentaires s'associent et se dissocient réversiblement suivant les principes thermodynamiques (action de la chaleur) et/ou d'ionisation du milieu (salinité du milieu). Les puces à ADN permettent de détecter et de quantifier simultanément l'expression de plusieurs milliers de gènes dans une cellule ou un tissu, à un instant précis, ayant subit des traitements différents.

Le principe des puces à ADN

Une puce à ADN est un support solide (verre ou nylon) de quelques centimètres carrés, sur lequel sont déposées quelques milliers de spots. Un spot correspond à une sonde, chaque sonde représente de nombreuses copies d'une séquence d'ADN spécifique d'un gène ou d'une famille de gènes donnée.

Comme pour les techniques de Southern et de Northern, les sondes peuvent être des oligonucléotides (simples brins) de synthèse, des produits de PCR ou des gènes exprimés (Expressed Sequence Tags) de séquence connue et spécifique d'un gène. Les sondes ont pour rôle de détecter des cibles marquées complémentaires, présentes dans le mélange complexe à analyser (ARNm extrait et converti en ADNc). La séquence d'ADNc cible est marquée et préparée en solution et mise en contact avec la puce à ADN portant les sondes.

Après acquisition des images d'hybridation, les signaux d'hybridation sont détectés selon le type de marquage, radioactivité ou fluorescence, et quantifiés. La quantification reflète le niveau d'expression, de chacun des gènes ou de la famille de gène représentés sur la puce.

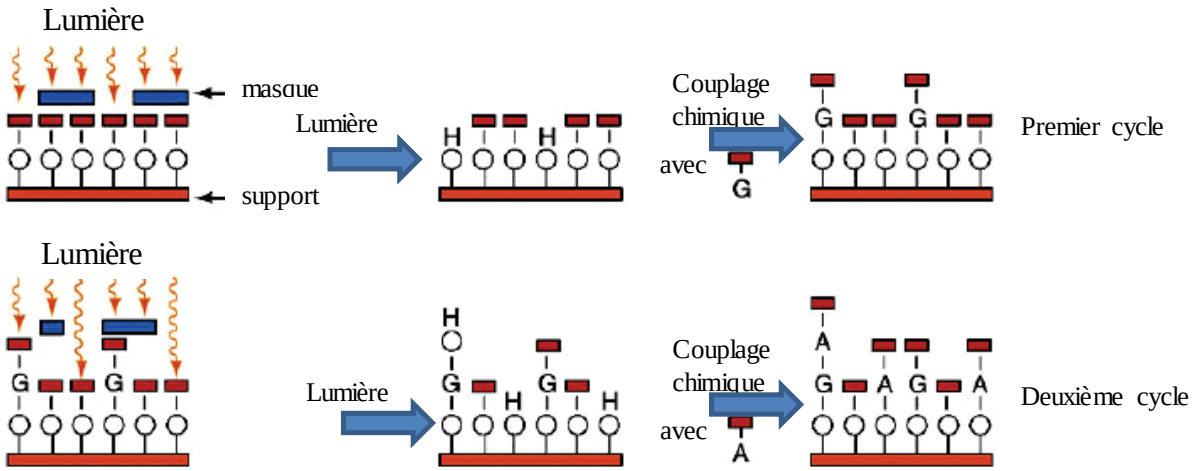


Figure 20: Synthèse du masque lithographique selon la technologie Affymetrix.

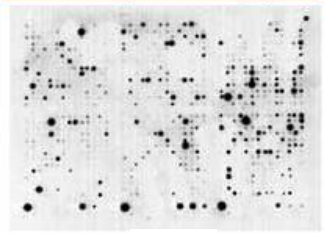
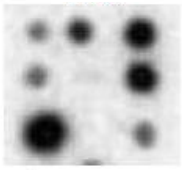
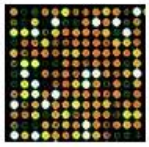
	Filtres de hautes densité macroarrays	Support en verre microarrays	Puces à oligonucléotides
	 détail 	 détail 	 détail 
Taille	18 cm x 8 cm	5.4 cm x 0.9 cm	1.28 cm x 1.28 cm
Densité de clones	2400	10 000	300 000
Type de marquage	radioactif	fluorescent	fluorescent
Nombre de conditions	1	2	1

Figure 21: Comparaison des puces à ADNc et des puces à oligonucléotides.

D'après (Lecrom and Marc, 2011).

Le support

Le support solide sur lequel sont fixées les sondes est une surface (matrice) inférieure à $1cm^2$, plane ou poreuse (percées de puits), composée de matériaux tels que le verre, les polymères, le silicium ou l'or et le platine. L'unité d'hybridation de la puce est le plot, élément principal de la puce à ADN, sur lequel est fixée la sonde d'ADN. Les plots sont répartis régulièrement sur toute la surface de la puce.

Les sondes

Les sondes sont fixées sur le support par fixation d'oligonucléotides ou par synthèse *in situ*.

La fixation d'oligonucléotides synthétisés permet de créer des sondes relativement longues, jusqu'à 40 à 60 bases, par adressage mécanique, grâce à une micropipette robotisée, ou par adressage électrochimique par activation d'une microélectrode spécifique sur laquelle est réalisée une éléctro-polymérisation.

La fixation d'oligonucléotides par synthèse *in situ* est effectuée par dépôt de couches successives de nucléotides A, T, G, C sur le support en verre (Fig. 20). Un masque lithographique permet l'empilement correct de 30 bases au maximum. L'ajout de base à la sonde se fait par adressage photochimique. Ce type de fixation est utilisé dans les puces Affymetrix.

A l'hybridation, chaque sonde reconnait, dans le mélange appliqué à la surface de la puce, la séquence d'ADN cible qui lui est complémentaire. Cette phase d'hybridation est suivie d'un lavage destiné à débarrasser la puce des cibles nucléiques non hybridées. Les hybridations sont détectées par lecture optique, les sondes hybridées avec les cibles marquées sont alors révélées.

Il existe deux types de puces, les puces à ADNc et les puces à oligonucléotides (Fig. 21). Parmi les puces à ADNc, on compte les macroarrays et les microarrays. Les puces à ADN microarray et macroarray diffèrent selon la densité de sondes et le type de détection des signaux d'hybridation: les macroarrays sont à faible densité de spots et le signal d'hybridation sonde-cible est détecté par radioactivité, les microarrays présentent une forte densité de spots et la détection du signal d'hybridation se fait par fluorescence.

1.3.1.2.1.1. Puces à ADNc

1.3.1.2.1.1.1. Macroarray

Les produits PCR sont déposés sur membranes de nylon à faible densité: environ 40 dépôts/cm^2. Les transcrits, à analyser, sont issus d'une seule condition expérimentale. Ces cibles sont rétro-transcrits en ADNc, puis marqués au phosphate radioactif ^{33}P ou ^{32}P.

Ces cibles radioactives sont hybridées avec les sondes sur la membrane de nylon. Les signaux d'hybridation radioactifs sont détectés et révélés par un film radiographique ou un appareil de détection de rayon X, puis quantifiés et analysés à l'aide de logiciels d'analyse d'images.

1.3.1.2.1.1.2. Microarray

Les produits PCR ou oligonucléotides longs (50mer-70mer) sont déposés sur lames de verre, à plus forte densité (comparé aux macroarray) : jusqu'à 6000 dépôts / cm^2.
Les ARNm sont extraits de deux échantillons à comparer, puis rétro-transcrits en ADNc.
Pour chaque échantillon, les ADNc sont marqués spécifiquement par une molécule fluorescente. Couramment, l'un est marqué à la cyanine 3 (Cy 3, fluorochrome vert), l'autre à la cyanine 5 (Cy 5, fluorochrome rouge). Les Cy3 et Cy5 sont des marqueurs fluorescents solubles dans l'eau, permettant la visualisation et la quantification d'acides nucléiques et de protéines. L'utilisation de deux fluorochromes différents permet la détermination des signaux d'hybridation de deux échantillons distincts au cours d'une seule expérience. Afin d'éliminer les biais résiduels, un marquage en « dye swap » consiste à réaliser deux hybridations en inversant les marquages des fluorochromes.

L'hybridation des cibles sur les sondes est compétitive: plus la concentration d'une cible est élevée, plus cette cible s'hybride sur la sonde. La couleur rouge ou verte de l'intensité de fluorescence revèle une hybridation majoritaire d'un des échantillons d'ADNc.

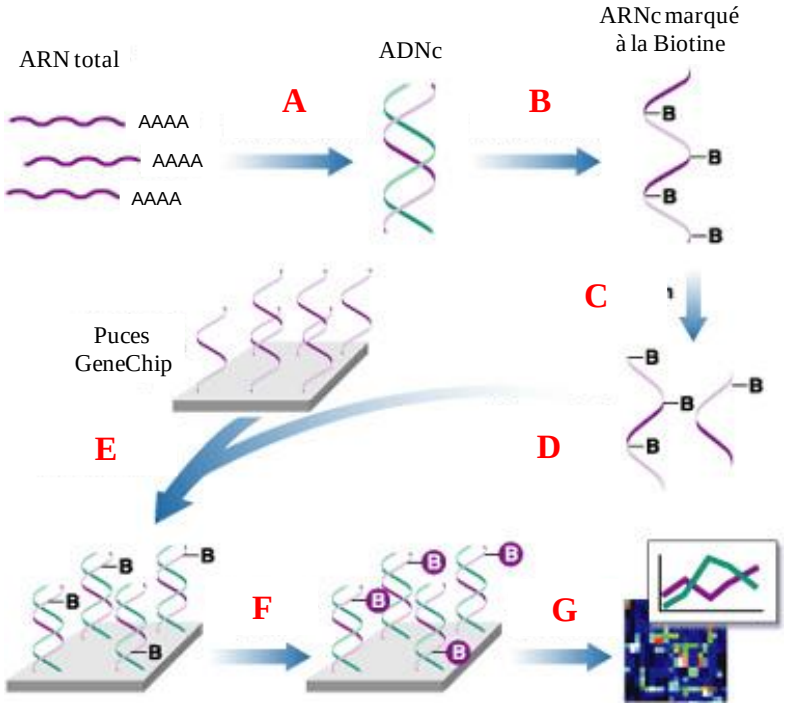


Figure 22: Schéma des différentes étapes de la technologie Genechips Affymetrix.

(A) Extraction des ARNs de l'échantillon, (B) Transcription inverse des ARNs, (C) Amplification et marquage (biotine) des ADNc, (D) Fragmentations des produits marqués et amplifiés, (E) Hybridation des amplifias sur la puce, (F) Elimination par lavage des produits non fixés, (G) Quantification, analyse des produits fixés et correpondance avec les ARNs de l'échantillon de départ soit des gènes exprimés dans la condition. D'après http://www.dkfz.de/gpcf/24.html.

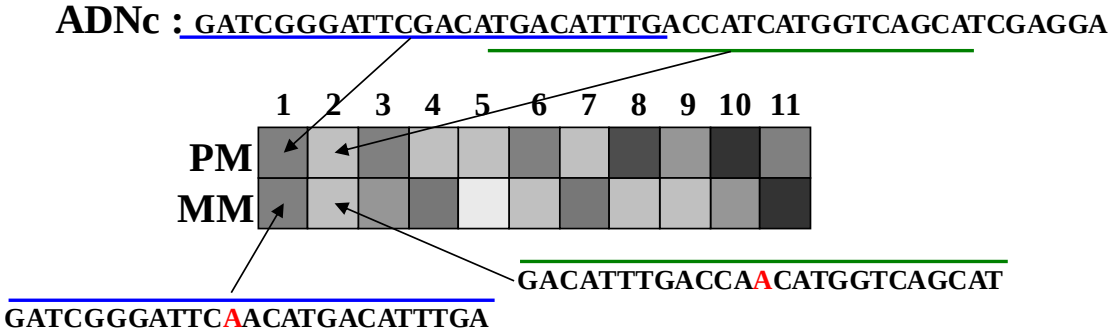


Figure 23: Schéma d'un probeset de la technologie Affymetrix.

Une puce est un support solide composé de milliers d'emplacements carrés, chacun contient des millions de copies d'un oligonucléotide La mesure de l'expression d'un gène utilise, ici, 11 paires de sondes (carrés sur le schéma), les 11 paires représentent 1 probeset, soit 11 sondes PM (Perfect Match) et 11 sondes MM (MisMatch).

1.3.1.2.1.2. Puces à oligonucléotides

Les puces à oligonucléotides sont une technologie dominante par rapport aux puces à ADNc. Parmis ces puces à oligonucléotides, on retrouve les technologies Affymetrix, Agilent, Nimblegen, et Spotted arrays.

La technologie Genechip Microarray Affymetrix (Fig. 22) utilise des oligonucléotides de 25pb synthétisés *in situ* sur une matrice de verre, et fixés par photolithographie. Toutes les séquences incluses sur la matrice sont sélectionnées à partir de GenBank, dbEST et RefSeq. Chaque gène est représenté par 11 à 20 sondes décrivant un probeset (une couverture de différentes régions de la séquence du gène plutôt en 3'). La technologie Wheat GeneChip Genome Array, utilisée dans mes travaux de thèse, comporte 11 sondes (Fig. 23).

Afin d'identifier d'éventuelles hybridations croisées et d'éliminer le bruit de fond résultant, chaque probeset est représenté par deux types d'oligonucléotides: des oligonucléotides parfaitement complémentaire à la séquence du gène de référence (« Perfect Match » ou PM) et des oligonucléotides présentant une erreur sur la base du milieu, typiquement la 13ème (Mismatch ou MM). À chaque sonde PM est associée une sonde MM. Les MM correspondent à l'hybridation non spécifique. Le procédé de préparation des ARNs cibles d'un échantillon se déroule en différentes étapes (Fig. 22): l'ARN est transcrit en ADNc double brin, puis en ARNc par transcription *in vitro* et marqué à la biotine pour l'hybridation. Après lavage l'ARNc marqué et hybridé à la sonde est révélé par coloration grâce à la streptavidine-phycoérythrine. Les biopuces sont analysées avec un module de criblage GeneArray à une longueur d'onde de 488 nm. L'émission de lumière à une longueur d'onde de 570 nm est proportionnelle à la quantité de cible liée à chaque position des oligonucléotides sur la puce.

La technologie NimbleGen utilise des oligonucléotides de 50-75 nt, synthétisés *in situ* par procédé photochimique utilisant un masque électronique (DLP), et compte 2.1x10^{6}sondes/puce.

La technologie Agilent utilise des oligonucléotides de 25-60 pb, synthétisés *in situ* et compte 244x10^{3} sondes/puce. Le dépôt des sondes sur la lame se fait de manière similaire à celle de l'impression à jet d'encre. Les multiples pointes des robots spotteurs déposent des rangées d'infimes gouttelettes d'oligonucléotides à des positions spécifiques de la puce.

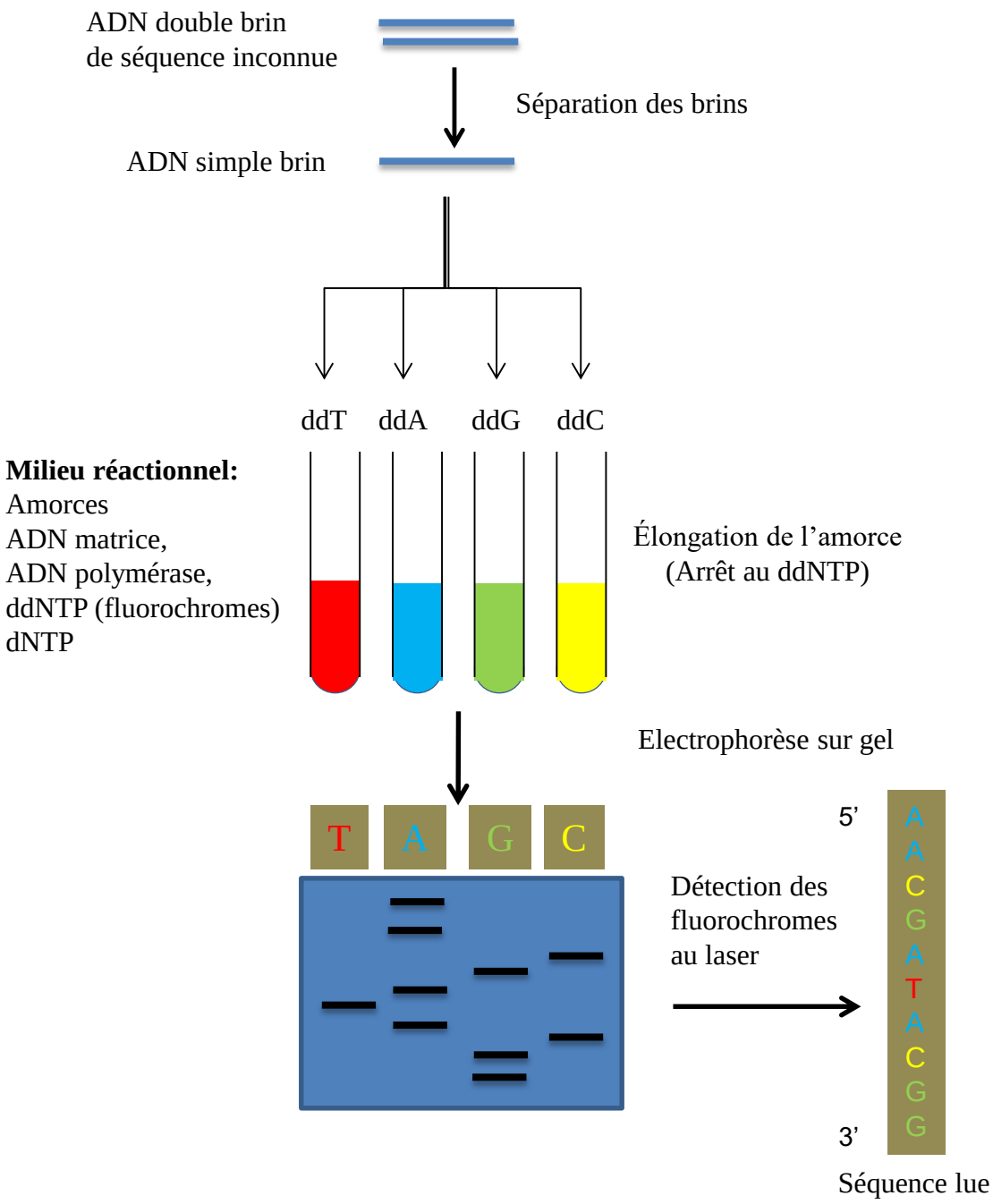


Figure 24: Schéma du séquençage par la méthode de Sanger.

L'amorce se lie à l'ADN à séquencer. La fixation des ddNTP génère des brins partiels. Les produits de chaque réaction déposés sur le gel d'électrophorèse se séparent suivant leurs tailles créant un profil en échelle (le plus fragment petit migre le plus loin). L'excitation du fluorochrome au laser permet de détecter le nucléotide en dernière position et détermine l'enchaînement des nucléotides en remontant les profils (en échelle) des quatre pistes simultanément.

1.3.1.2.2. Le Séquençage

Le séquençage détermine l'enchaînement des nucléotides (Adénine, Cytosine, Thymine, Guanine) sur un brin d'ADN, qui constituent l'information génétique. Historiquement, le séquençage a été développé dans les années 1970, par deux équipes indépendantes : Maxam et Gilbert (Maxam and Gilbert, 1977) aux Etats-Unis et Sanger (Sanger et al., 1977) en Angleterre, proposant deux méthodes différentes. L'une est fondée sur la synthèse enzymatique (méthode de Sanger), l'autre sur la dégradation chimique (méthode de Maxam et Gilbert).

La première génération de séquençage réunit toutes les technologies fondées sur ces deux méthodes, tels que le séquençage par mesure de la fluorescence et par électrophorèse sur gel acrylamide ou capillaire qui sont apparues vers les années 80 et 90 respectivement.

1.3.1.2.2.1. Les premières méthodes de séquençage

1.3.1.2.2.1.1. La méthode de Sanger

Le principe consiste à synthètiser en parallèle des copies partielles et intermédiaires du fragment d'ADN initiale à séquencer.

La première étape est une polymérisation, de l'ADN à séquencer, initiée à partir d'une amorce complémentaire. L'activité d'élongation se fait par le fragment de Klenow (une ADN polymérase I dépourvue d'activité exonucléase dans le sens 5'→3') utilisant les quatre désoxyribonucléotides (dATP, dCTP, dGTP, dTTP) nécessaires et une plus faible quantité d'un des quatre didésoxyribonucléotides[7] marqués (radioactifs) (ddATP, ddCTP, ddGTP ou ddTTP). L'incorporation aléatoire du didésoxyribonucléotide (ddNTP), dans le brin en cours de synthèse, arrête l'élongation.

Le séquençage d'un même fragment d'ADN nécessite 4 réactions en parallèle, chacune menée avec un seul des quatre didésoxyribonucléotides. Dans chaque série, l'utilisation d'un ddNTP permet d'obtenir un ensemble de fragments intermédiaires d'ADN (typiquement de 650–800 bp), arrêté à l'incorporation du ddNTP (Fig. 24).

Les fragments de chaque série sont ensuite séparés sur quatre pistes différentes, selon leur taille par migration par électrophorèse sur gel de polyacrylamide.

[7] Les ddNTP diffèrent des dNTP par l'absence d'un groupement OH en position 3'. Lorsque l'ADN polymérase utilise un ddNTP, la synthèse du brin d'ADN s'arrête.

À la lecture des quatre pistes simultanément, on reconstitue base par base l'enchaînement des nucléotides du fragment d'ADN initiale. Suivant la taille des gels, la séquence est limitée à 1kb maximum en 6-8 heures, avec une lecture par échantillon.

Des améliorations de cette technique de séquençage ont remplacé :

- les didésoxynucléotides radioactifs par un marquage avec différents fluorophores permettant les quatre réactions dans le même tube,
- -et la lecture de la séquence par un séquençage automatisé des produits de PCR par électrophorèse à haute résolution.

Ces améliorations ont abouti à l'apparition de séquenceurs automatiques (dans les années 1990), permettant de séquencer un grand nombre d'échantillons en parallèle et une diminution du coût.

La méthode de Sanger a été utilisée pendant plus de 30 ans. Cette technologie a été utilisée, notamment pour le séquençage entier du génome humain en 2003. La réalisation de ce projet représente 13 ans d'efforts et un coût estimé à 2.7 milliards de Dollars (Lander et al., 2001; IHGSC, 2004 ; Cole et al., 2008), par la suite d'autres projets ont été initié dont le séquençage des génomes d'*Arabidopsis* (Initiative, 2000) et du riz (Eckardt, 2000).

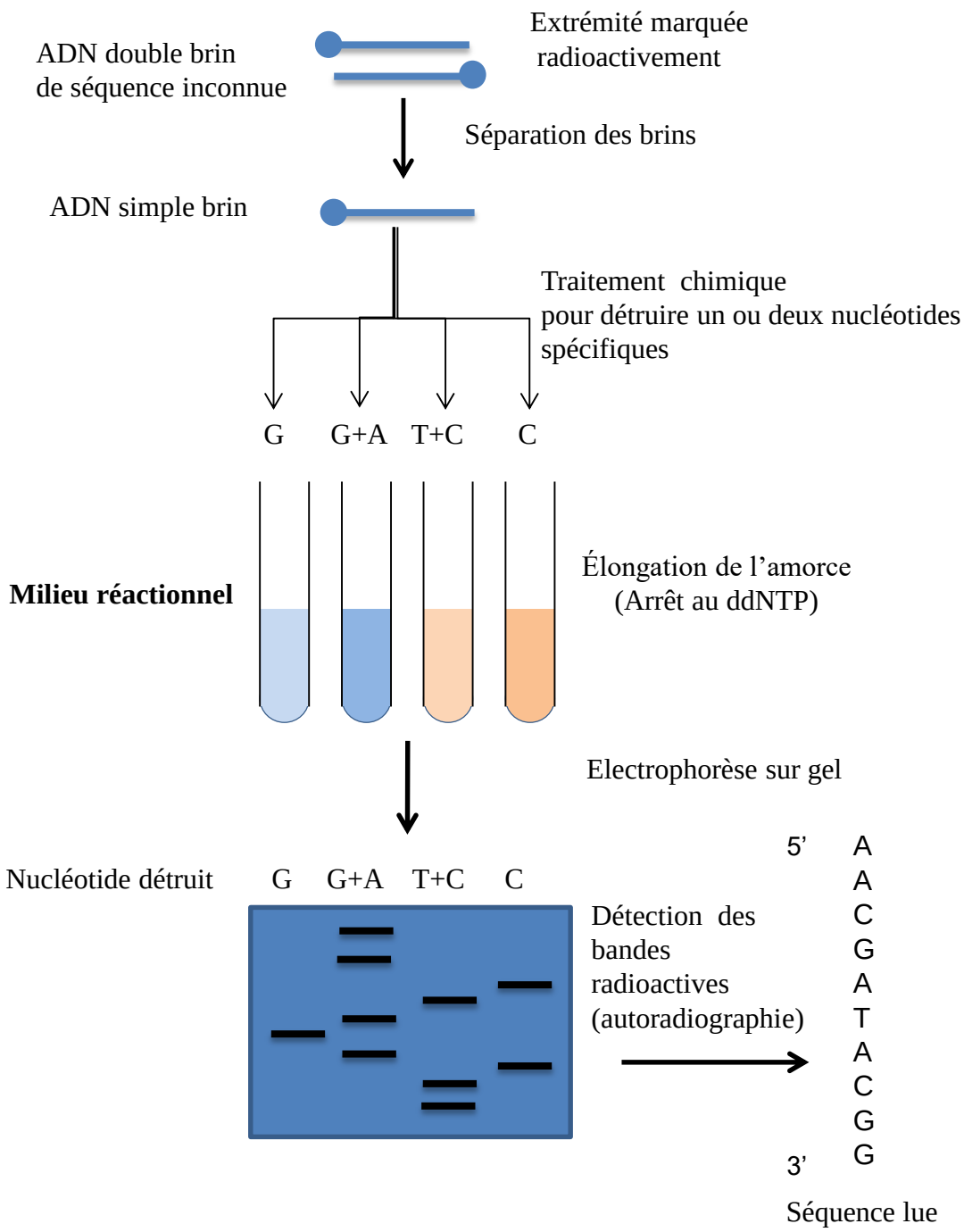


Figure 25: Séquençage par la méthode de Maxam-Gilbert.

Cette méthode utilise des réactifs qui détruisent des bases nucléotidiques spécifiques, clivant ainsi la molécule d'ADN en des sites spécifiques.

1.3.1.2.2.1.2. la méthode de Maxam-Gilbert

Cette méthode chimique a pour principe de déterminer l'enchaînement des nucléotides par cassure chimique et différentielle (Fig. 25). L'ADN à séquencer est marqué radioactivement, à l'extrémité 5', par le P^{32}, dénaturé puis purifié sur gel de polyacrylamide.

Le séquençage nécessite, là aussi, la réalisation de 4 réactions différentes sur un même échantillon.

A chaque tube correspond un traitement chimique, permettant de détruire un nucléotide spécifique (A, C, G ou T) et donc de rompre le fragment d'ADN spécifiquement à ce site. Les différents traitements distinguent dans un premier temps les purines et les pyrimidines.

Pour les purines, le diméthysulfate accompagné d'une augmentation de la température coupe l'ADN au niveau des Guanines, tandis qu'additionné d'un traitement acide, il coupe au niveau des deux bases (Adénines et Guanines).

Pour les pyrimidines, l'hydrazine accompagné d'une forte concentration en chlorure de sodium (NaCl) coupe au niveau des Cytosines, tandis qu'employé seul il coupe au niveau des pyrimidines (Cytosines et Thymidines).

Les fragments produits sont séparés par électrophorèse en gel de polyacrylamide, pour révéler l'enchaînement des nucléotides de la séquence tel que décrit précédemment.

1.3.1.2.2.2. Les Nouvelles Technologies de Séquençage

En 2007, est apparu le séquençage de nouvelle génération (ou NGS: Next Generation Sequencing) permettant de générer des débits jusqu'a 1000 fois supérieurs à ceux obtenus par le séquençage classique et d'éviter certains biais de la méthode de Sanger, dus notamment au clonage de l'ADN à séquencer.

La méthode RNA-sequencing consiste à séquencer des ARNm (ou tout ARN issu du transcriptome) converti en ADNc. Il permet d'annoter des génomes, de repérer les gènes exprimés, d'identifier de nouveaux gènes mais également de quantifier l'expression des gènes en comptant le nombre de transcrits séquencés pour un gène donné.

Actuellement, plusieurs technologies sont en concurrence sur le marché, développant différentes méthodes de séquençage par synthèse ou pyroséquençage qui génèrent des fragments de tailles différentes variant de 100 à 1000 pb selon la technologie employée. Je détaille, ici, les principales méthodes utilisées:

- le pyroséquençage,
- le séquençage par synthèse,
- le séquençage par nano-mesure de pH,
- le séquençage par ligation,
- le séquençage « single molecule ».

Les plateformes NGS partagent une caractéristique technologique commune: le séquençage en parallèle massif de clones amplifiés ou de molécules d'ADN simple brin, séparées physiquement dans une cellule d'écoulement ou « flow-cell ». Cette méthode diffère du séquençage de Sanger, basé sur la séparation par éléctrophorèse des copies intermédiaires produites dans différentes réactions de polymérisation partielles (présenté dans le paragraphe précédant). Dans les NGS, le séquençage est démultiplié par des cycles répétés de polymérisation. Selon le procédé de la plateforme, les NGS génèrent des centaines de mégabases voir de gigabases de séquences nucléotidiques lues pour un seul échantillon dans un fichier unique de sortie pour un même processus ou « run »[8] (Voelkerding et al., 2009).

[8] Un run (réalisation d'un processus complet par la machine) produit un grand nombre de lectures (reads) correspondant à des séquences d'ADN ou d'ARN de l'espèce étudiée.

Ce nombre important de séquences permet par comptage du nombre de lecture pour un gène donné de quantifier son état d'expression. Les technologies des diverses plateformes se différencient par la nature des réactifs, des substrats utilisés mais également par le mode de détection d'incorporation des bases.

Lecture simple et lectures pairées

Les méthodes NGS permettent de séquencer en parallèle des millions de fragments d'ADN, mais ces lectures sont courtes comparées à celles produites par la méthode de Sanger. Ces lectures de petites longueurs sont une limite pour les résultats finaux de séquençage.

Le séquençage peut être fait à partir de l'une des deux extrémités (« single-end sequencing ») ou des deux extrémités du même brin d'ADNc (« paired-end sequencing »).

Le séquençage des lectures pairées est une stratégie pour améliorer les performances d'assemblage des séquences d'ADN en doublant l'information de séquence. Ces lectures sont séparées par une distance connue et lié à la technologie utilisée. Les étiquettes des extrémités en 5′ et 3′ d'un même fragment d'ADNc sont éloignées de centaines de pb (<600pb au maximum en moyenne). En tenant compte de cette distance l'assemblage des lectures permet des vérifications par recouvrement.

1.3.1.2.2.2.1. Le pyroséquençage

La technologie 454 (http://www.454.com) dérive du pyroséquençage et de la PCR en émulsion.

Le pyroséquençage est une approche basée sur la détection par chimiluminescence du groupement pyrophosphate libéré lors de l'incorporation d'un dNTP (déoxynucleotide triphosphate) par la polymérase. Cette approche a été développée par Nyren et al, (Nyren et al., 1993), puis améliorée (Ronaghi et al., 1996; Ronaghi et al., 1998) pour devenir la base du pyroséquençage. Durant la même période, Tawfik et Griffiths (Tawfik and Griffiths, 1998) ont développé la PCR en molécule unique, dans des micro-compartiments en émulsion oléo-aqueuse. Cette méthode porte également le nom de « PCR en émulsion ».

En 2000, J.Rothberg fonde 454 Life Sciences, qui développent la première plate-forme NGS disponible dans le commerce : le GS20 est lancé en 2005. Margulies et al (Margulies et al., 2005) combinent la PCR en émulsion et le pyroséquençage pour effectuer le séquençage en masse

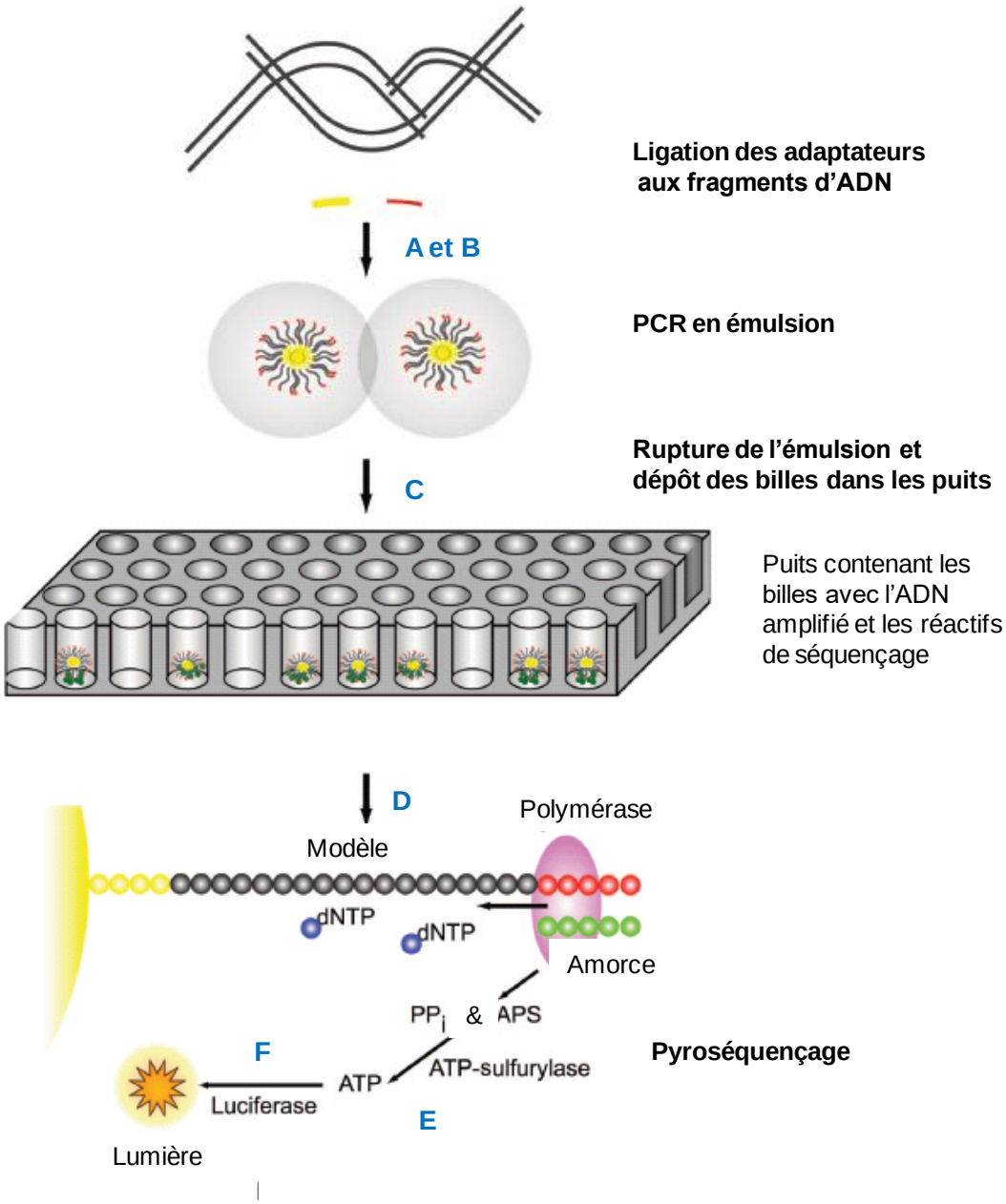


Figure 26: Séquençage Roche 454 GS FLX.

L'échantillon d'ADN (A) est fragmenté et lié aux adaptateurs, (B) Amplification des fragments par PCR en émulsion (sur billes), (C) Répartition des billes de l'émulsion dans les micro-puits en présence des réactifs de séquençage, (D) Séquençage itératif avec la libération d'un pyrophosphrate (PPi) lors de l'intégration d'un nucléotide au cours de la polymérisation, (E) Sous l'action enzymatique de l'ATP-sulfurylase en présence de PPi et d'APS (adenosine 5'-phosphosulfate) il y a formation d'ATP. (F) L'enzyme luciférase induit une réaction entre la luciférine et l'ATP, en émettant une lumière mesurable et témoin de l'incorporation d'un nucléotide marqué. La séquence reconstituée correspond à l'alternance des informations mesurées entre l'ordre de passage des nucléotides marqués avec un fluorophore et l'incorporation de celui-ci avec l'émission de lumière pour chaque emplacement, d'après (Voelkerding et al., 2009).

ou « shotgun » et l'assemblage *de novo* du génome de *Mycoplasma genitalia*[9] avec une couverture de 96% et une précision de 99.96% en un seul processus du GS 20.
En 2007, Roche Applied Science acquière 454 Life Sciences, et introduit une nouvelle version de l'outil 454 : le GS FLX, amélioré aujourd'hui en GS FLX Titanium XL+.

La bibliothèque d'ADN matrice est préparée par fragmentation (nébulisation ou sonication), puis des adaptateurs d'oligonucléotides sont liés aux extrémités des fragments de plusieurs centaines de paires de base (Fig. 26). La bibliothèque est ensuite diluée à la concentration d'une seule molécule, dénaturée, puis hybridée sur des billes individuelles contenant des séquences complémentaires aux oligonucléotides adaptateurs. Les billes sont compartimentées en microvésicules dans lesquelles a lieu la PCR. Après l'amplification du nombre de copies, la PCR en émulsion est interrompue.

Les billes (microréacteurs) enrichies en nombre de copies sont de nouveau séparées par dilution limitante, déposées dans les puits individuels de réactions de séquençage (de l'ordre de 106 puits, à l'échelle du pico-litre, 10^{-12}), en présence des enzymes de séquençage, et de billes supplémentaires couplées de deux enzymes : une sulphurylase et une luciférase.

La réaction de synthèse commence par la fixation d'une amorce sur l'adaptateur puis chaque base dNTP est ajoutée séquentiellement pour compléter le fragment à séquencer, à partir de l'extrémité 3' de l'amorce. Chaque base est marquée par un fluorophore différent et l'incorporation d'un nucléotide induit la libération d'un pyrophosphate (d'où le terme «pyroséquençage»). L'ATP-sulfurylase transforme ce pyrophosphate en ATP qui est ensuite utilisé et couplé à une luciférine par une luciférase. La réaction libère une oxyluciférine et un signal lumineux. Cette luminescence est capturée par une caméra CCD. À chaque flux de dNTP, les puits sont imagés, analysés pour un ratio signal/bruit, filtrés selon les critères de qualités, et traduits en une séquence de nucléotides.

Une caractéristique importante de la technologie 454 est la longueur des lectures, qui facilite l'assemblage *de novo* des génomes (Pearson et al., 2007). Malgré un coût élévé, cette technologie est avantagée par la longueur des lectures ou « reads », qui a grandement évolué, de 100pb/run en 2006 à 1kb/run aujourd'hui, généré en 24h (avec 0.7Gb de données en sortie par run)(Liu et al., 2012b).

[9] bactérie, agent infectieux pathogène pour l'homme, responsable d'urétrites et d'autres maladies sexuellement transmissibles

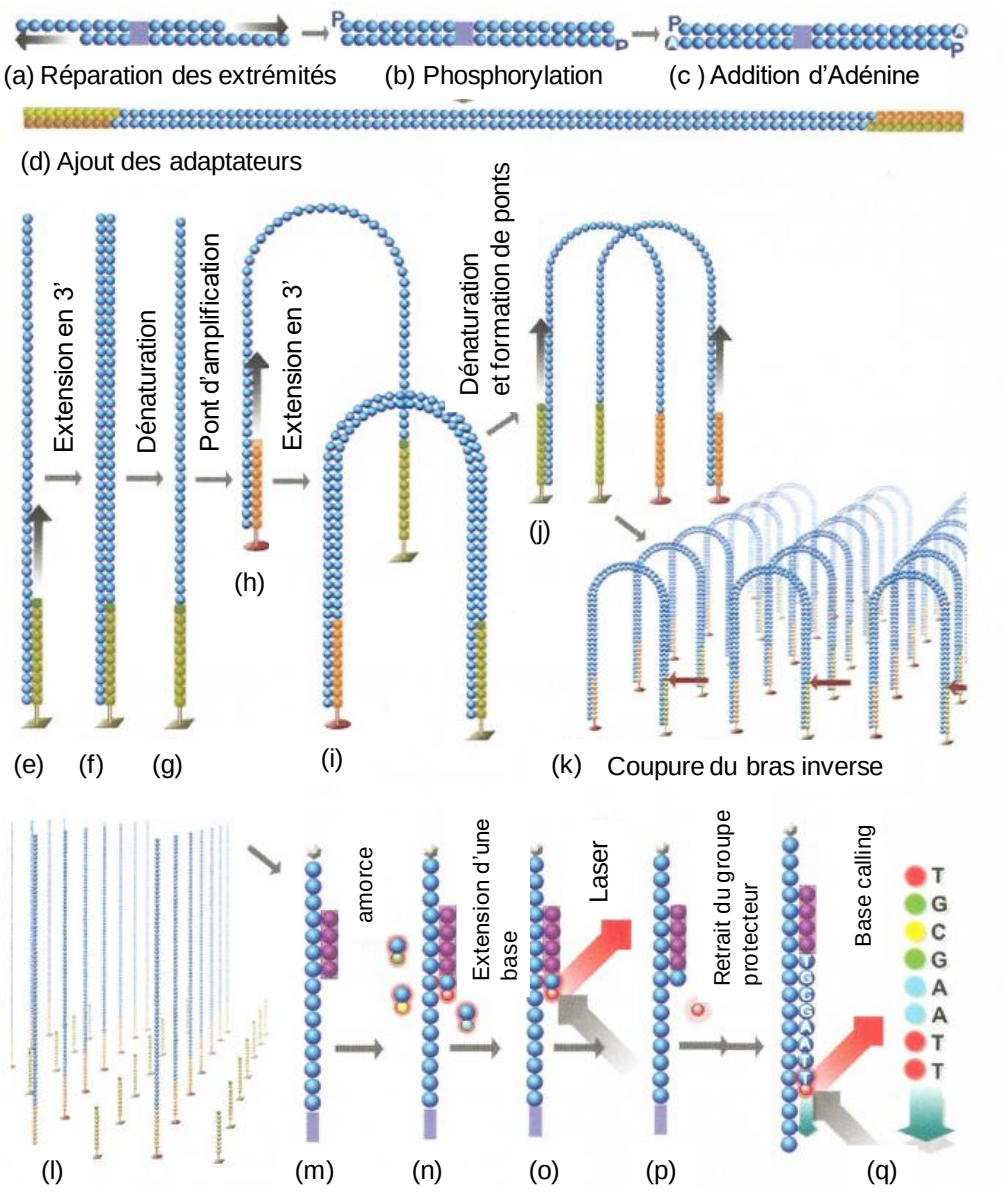


Figure 27: Le séquençage par synthèse.

Préparation de la librairie :a)les fragments d'ADN sont réparés, b) phosphorylés, c) l'Adénine y est ajoutée, d) des adaptateurs sont liés aux extrémités. suivi par la formation de clusters : e) l'ADN dénaturé est lié aux amorces sur la surface de la cellule d'écoulement ou « flow cell », f) le brin complémentaire est synthétisé, g) puis dénaturé. h) L'ADN simple brin forme un pont avec l'amorce adjacente sur la surface du flow cell, i) puis est étendu de nouveau pour créer un pont double brin, j) qui est de nouveau dénaturé et s'hybride à d'autres amorces pour former des ponts simple brin. k) Le processus est répété 35 fois, pour former des clusters d'environ 2000 molécules. Le brin reverse est clivé (flèche marron). L'extrémité 3' est bloquée et les brins sont hybridés aux amorces de séquençage m). Le séquençage : l'ADN polymérase incorpore un terminateur réversible n), la fluorescence est enregistrée o), l'étiquette fluorescente est éliminée et le produit est libéré p). Le processus est répété plus de 50 cycles pour déterminer la séquence (d'après www.Illumina.com).

1.3.1.2.2.2.2. Le séquençage par synthèse

En 1997, les chimistes anglais Shankar Balasubramanian et David Klenerman conceptualisent une approche de séquençage de molécules d'ADN simple attachées à des microsphères. Ils fondent Solexa en 1998. Leurs objectifs pendant les premiers développements de séquençage de molécule d'ADN n'ont pas été atteints, exigeant un changement vers le séquençage de clones amplifiés. En 2006, le Genome Analyseur de Solexa, la première plateforme de séquençage de petites lectures, est commercialement lancée. Acquis par Illumina (http://www.Illumina.com) en 2006, le Genome Analyseur utilise une cellule d'écoulement ou « flow-cell », qui est une surface plane transparente semblable à une lame de microscope, constituée de 8 lignes individuelles sur lesquelles des oligonucléotides sont ancrés à la surface, permettant l'hybridation des librairies (banques d'ADN) grâce aux adaptateurs (Fig. 27).

Les échantillons d'ADN sont découpés en fragments de plusieurs centaines de paires de bases et réparés aux extrémités pour générer une extrémité franche en 5'. L'activité polymérase du fragment de Klenow permet d'ajouter une base A à l'extrémité 3'. Cet ajout prépare les fragments d'ADN pour la ligation aux oligonucléotides adaptateurs, composés d'une succession de base T à l'extrémité 3', pour augmenter l'efficacité de liaison. Les oligonucléotides adaptateurs sont complémentaires à ceux ancrés sur la cellule d'écoulement (flow-cell). Dans des conditions limitantes de dilution, les molécules d'ADN uniques (simple brin) liées aux adaptateurs sont déposées et immobilisées aléatoirement par hybridation aux oligonucléotides ancrés sur le flow-cell.

Une fois la bibliothèque hybridée à l'intérieur du flow-cell, le brin complémentaire d'ADNc est synthétisé par les polymérases. L'ADN/ADNc double brin est dénaturée et la molécule d'ADN simple brin originale est éliminée. La molécule d'ADNc nouvellement synthétisée forme une liaison covalente aléatoirement avec l'oligonucléotide (ou primer) adjacent présent à la surface du flow-cell, pour former un pont. Le primer hybridant l'oligonucléotide sur le flow cell permet la synthèse du brin d'ADNc inverse grâce aux polymérases. Puis, le pont d'amplification est dénaturé. Le cycle du pont d'amplification est répété jusqu'à la formation de nombreux ponts.

Ces ponts sont alors dénaturés, les brins réverses sont clivés et éliminés afin de laisser que le brin forward pour un séquençage en sens unique. L'amplification des fragments d'ADN déposés crée jusqu'à 1000 copies identiques de chaque molécule, appelé des clusters, chacun d'eux étant unique. Cette amplification aboutit à la formation de 600-800K clusters/mm². Le flowcell est alors prêt pour le séquençage.

Le séquençage du brin forward est initié par hybridation d'une amorce complémentaire avec des adaptateurs, suivi de l'ajout d'une polymérase et d'un mélange des 4 dNTP portant un colorant fluorescent réversible (une couleur par nucléotide). Ces nucléotides sont incorporés par complémentarité de séquence, dans chaque brin du cluster. À ce stade, le colorant fluorescent réversible bloque la polymérisation. Après incorporation, les réactifs en excès sont éliminés. Puis, après une excitation au laser, chaque cluster sur le flow-cell émet une fluorescence, qui est capturée et enregistrée dans une image. Plusieurs réactions chimiques successives permettent de retirer le colorant fluorescent afin de débloquer la polymérisation et permettre l'incorporation du nucléotide suivant (Fig.27)

Ce séquençage par synthèse, itératif, nécessite 3-10 jours pour générer des lectures de 2x101pb de longueurs (1 base incorporée par cycle) (avec 600Gb de données en sortie par run) (Liu et al., 2012b).
L'acquisition des images se fait en temps réel à chaque cycle en fonction de chaque couleur émise. L'analyse d'images dit « base calling[10] », permet de déduire la séquence à partir de la détection de l'émission de fluorescence à chaque cycle.

La nouvelle plateforme, le Genome Analyseur II, bénéficie de modifications optiques permettant des analyses de densités plus élevées de cluster.

Une des limites du séquençage d'Illumina, outre la longueur des lectures très courtes, est la diminution de la qualité d'identification de base ou « base-calling » avec l'augmentation de la longueur des lectures (Dohm et al., 2008), ce qui est principalement lié au de déphasage.

[10] correspondance entre la fluorescence et la base pour un cluster

Le déphasage parasite les signaux émis par un cluster, en effet pendant un cycle de séquençage donné, les nucléotides peuvent-être sur ou sous-incorporés. Avec des cycles successifs, ces aberrations s'accumulent et produisent une population hétérogène dans un cluster avec des brins de longueurs différentes. Cette hétérogénéité diminue la pureté du signal et réduit la précision dans le base-calling, spécifiquement à l'extrémité 3' des lectures. Des améliorations dans la chimie de séquençage, les algorithmes d'analyse d'images et l'interprétation des résultats atténueraient le déphasage. Aussi, des améliorations dans la préparation de la bibliothèque, devraient augmenter la reproductibilité de la fragmentation (par sonication), dans l'objectif d'améliorer la ligation d'adaptateur, et de diminuer les biais G+C observés dans les lectures Illumina (Quail et al., 2008). Toutefois, cette technologie présente de nombreux avantages dont un rendement élevé et un faible coût.

1.3.1.2.2.2.3. Le séquençage par nano-mesure de pH

La technologie de séquençage Ion torrent de Life Technologies (Rothberg et al., 2011) utilise une combinaison unique de micro-fluidique et de technologie semi-conductrice permettant la transformation directe de l'information génétique en information numérique. Cette technologie se base sur les variations de pH pour détecter l'incorporation des nucléotides. Ce séquenceur de « détection de flux d'ions » comporte des biopuces composées de 165 et 660 millions de capillaires.

L'ADN simple brin est fixé sur des microbilles réparties au milieu des capillaires. Les nucléotides sont ajoutés séquentiellement pour la synthèse du brin complémentaire. Au cours de l'élongation, l'incorporation d'un nucléotide induit la libération d'un proton, qui entraine une variation du pH dans le capillaire. Cette modification du pH est détectée par la biopuce et enregistrée pour chaque capillaire.

Le « Personal Genome Machine » ou PGM génère 1 Gb de séquences par utilisation de puces de type 318, avec une taille moyenne de fragments séquencés d'environ 250 pb, en moins de 2 heures. L'Ion Torrent PGM 318 est compétitif dans cette catégorie, le MiSeq produit des séquences plus courtes (environ 150 pb) et le GS Junior présente un débit plus faible (35 Mb).

La technologie Ion Torrent présente l'avantage d'être peu coûteuse en réactifs notamment par l'absence de fluorochromes.

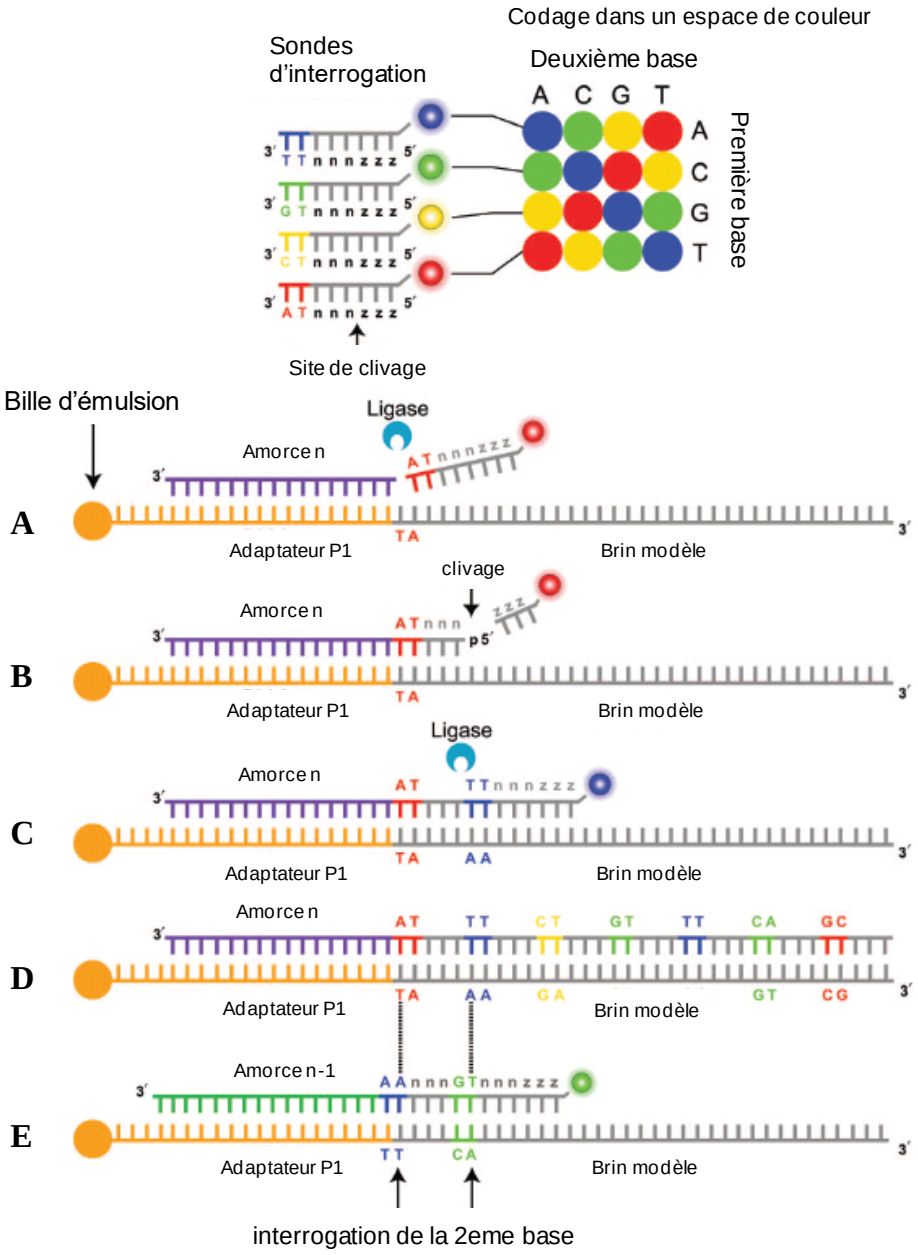


Figure 28: Applied Biosystems SOLiD séquençage par ligation.

En haut: le codage de l'espace couleur. Chaque sonde d'interrogation est un octamère, composé (dans le sens 3'-5') de 2 sondes spécifiques, suivies de 6 bases dégénérées (nnnzzz) et une des 4 étiquettes fluorescentes en 5'. Les 2 bases sondes spécifiques sont une des 16 combinaisons possibles. En bas : (A) adaptateur P1, brin modèle et amorce (n) hybridée. Le brin modèle est interrogé par une sonde, dont les 2 bases spécifiques et complémentaires du brin, ici, sont AT. (B) Après hybridation et ligation de la sonde, la fluorescence est enregistrée avant la coupure des 3 dernières bases dégénérées. L'extrémité 5' de la sonde clivée est phosphorylée avant l'étape suivante. (C) Hybridation de la seconde sonde. (D) Extension complète de l'amorce n à travers le premier "round" de 7 cycles. (E) L'extension de l'amorce (n) est dénaturée de l'adaptateur/brin modèle, et le second « round » de séquençage est effectué avec l'amorce (n-1), d'après (Voelkerding et al., 2009).

1.3.1.2.2.2.4. Le séquençage par ligation

La technologie SOLiD (Supported Oligonucleotide Ligation and Detection), a été initialement développée par la société Agencourt Personal Genomics (Beverly, MA, USA), puis acquise par Applied Biosystem en 2006 (http://www.solid.appliedbiosystems.com), qui a fusionné avec Life Technologies. SOLiD est une technologie de séquençage de petites lectures basée sur la ligation. Cette approche a été développée dans le laboratoire de George Church et décrite, en 2005, lors du re-séquençage du génome d'*Escherichia coli* (Shendure et al., 2005). Applied Biosystems a amélioré la technologie et a mis en vente l'instrumentation SOLiD en 2007.

La préparation de la bibliothèque est assez proche de la technologie 454, dans laquelle des adaptateurs sont liés aux fragments d'ADN, qui sont attachés à des billes et amplifiés par PCR en émulsion. Les billes portant les échantillons amplifiés sont immobilisées sur une plaque de verre (au lieu de micro-cuves comme pour la technologie 454). Le séquençage commence par l'hybridation de l'amorce à l'adaptateur (Fig. 28). Plutôt qu'une orientation dans le sens 5'→3' pour la polymérisation, l'amorce est orientée dans le sens 3'→5' pour la ligation des sondes d'interrogation pendant la première étape de séquençage.

Chaque sonde d'interrogation est un bloc de 8 bases, composé (dans le sens 3'→5') de 2 bases spécifiques suivies de 6 bases « dégénérées »[11] et portant une des 4 étiquettes fluorescentes en 5'. La fluorescence informe partiellement sur le couple de bases détectées. L'utilisation d'amorces de taille différente sur l'adaptateur (1 à 4 bases en moins par rapport à la première) permet d'identifier les bases et de les mesurer chacune deux fois (en décalé). Ce qui réduit le niveau des erreurs de lecture à 1 pour 1000, soit dix fois plus que dans le pyroséquençage. Les 2 bases spécifiques sont une des 16 combinaisons possibles (exemple : AT, TT, CT, ..) de bases complémentaires au brin à séquencer. Les 3 dernières bases dégénérées et l'étiquette fluorescente sont éliminées après ligation de la sonde.

[11] Les bases dégénérées ou universelles peuvent s'apparier indifféremment à l'une des 4 bases A,C,T,G, exemple : la déoxynosine (Guanine sans le groupement amine NH2)

Dans la première étape de séquençage, la ligase et les 16 combinaisons possibles de sondes d'interrogations sont présentes dans le milieu. Les sondes sont en concurrence pour s'hybrider sur la séquence d'ADN à séquencer. Après hybridation, un lavage retire les sondes non liées. Les signaux de fluorescence sont collectés avant le clivage des sondes liées, et un lavage permet de retirer l'étiquette fluorescente et de générer un groupement 5' Phosphate.
Dans les étapes suivant le séquençage, les sondes d'interrogation viennent se lier au groupement 5' Phosphate de la sonde précédente. Sept cycles de ligation, appelés « round », sont effectués pour la polymérisation de la première amorce. Le brin synthétisé est alors dénaturé, et une nouvelle amorce de séquençage diminuée d'une base dans sa séquence par rapport à la précédente (n-1) est hybridée sur l'adaptateur. Cinq rounds au total sont effectués, chaque round avec une nouvelle amorce de taille n-1 par rapport à la précédente. Dans ce processus chaque base est vérifiée deux fois grâce à la fixation alternée de deux amorces décalées de deux nucléotides permettant deux réactions de ligation indépendantes pour une même position. Le codage des résultats est effectué sur 2 bases dans un espace de 4 couleurs. La lecture des séquences est effectuée dans cet espace de couleur.

Le système de codage de la lecture sur deux bases permet une très grande fidélité de la lecture des résultats. Un processus de 7 jours génère des lectures de 50 bases (avec 120 Gb de données en sortie par run) (Liu et al., 2012b).

1.3.1.2.2.2.5. Le séquençage d'une unique molécule d'ADN « single-molecule »

HeliScope est la première plateforme de séquençage de troisième génération.
Le principe des séquenceurs de troisième génération repose sur le séquençage d'une molécule d'ADN sans amplification préalable par PCR, ce qui évite par conséquent les erreurs de lectures liées à la PCR ou les biais d'amplification vers les régions répétées (Pareek et al., 2011).

HeliScope a été developpée chez Helicos BioSciences (http://www.helicosbio.com) avec une sortie de séquençage de 1Gb/jour. Cette technologie de séquençage en direct, par synthèse d'une seule molécule d'ADN (technologie SMS : Single Molecule Sequencing), découle des travaux de Braslavsky en 2003 (Braslavsky et al., 2003).

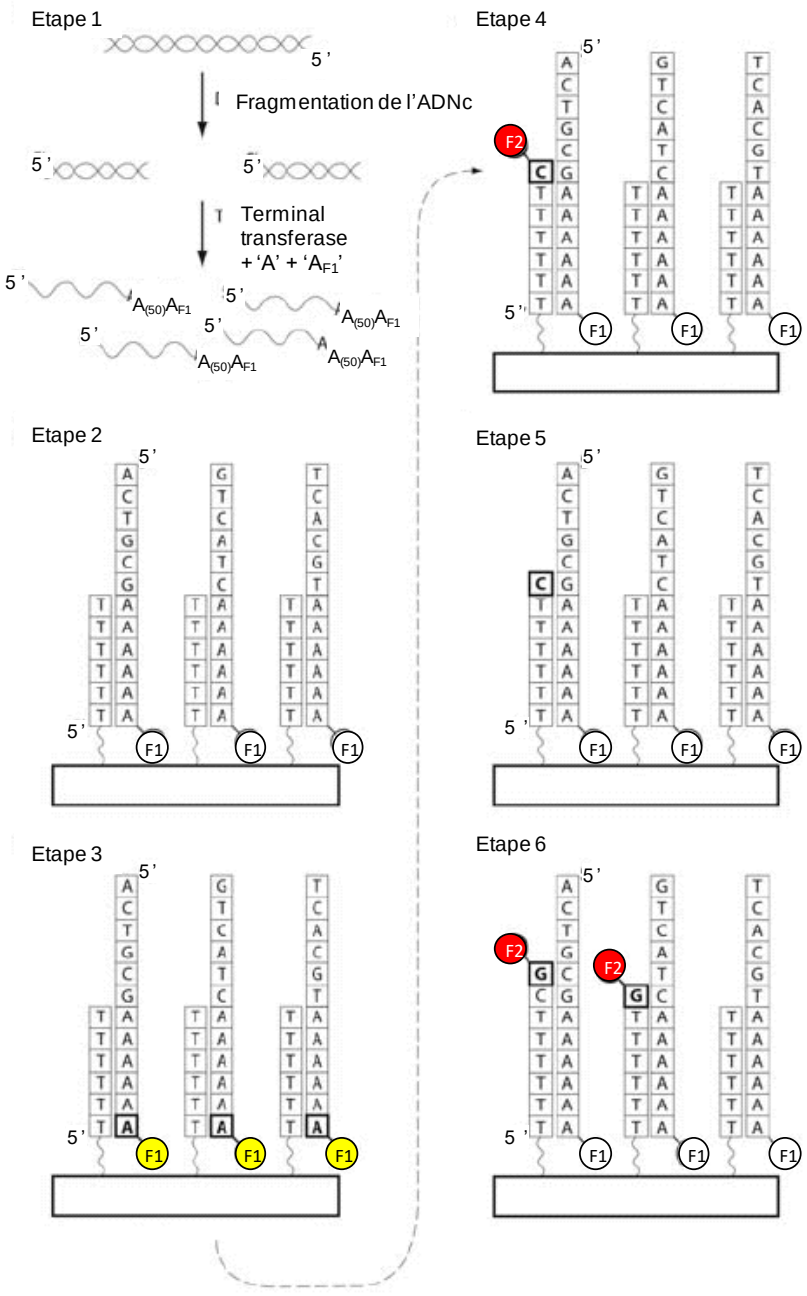


Figure 29: Illustration du séquençage "single-molecule sequencing"

(1)L'ADN génomique est préparé pour le séquençage par fragmentation et ligation par une transférase d'une queue polyA portant une étiquette fluorescente terminale. (2) Les séquences d'ADN sont hybridées sur la surface du flow cell par liaison covalente à une séquence poly(dT)(50 nucléotides). (3) Prise d'image de l'échantillon d'ADN pour établir les sites de fixation pour le séquençage par synthèse sur le support. (4) Elongation avec un nucléotide marqué par la polymérase, suivie d'un rinçage du mélange réactionnel et d'une capture d'image de l'étiquette de Cy5 excitée à 647nm. (5) Clivage de l'étiquette fluorescente et lavage. (6) Le cycle d'incorporation des autres nucléotides marqués continue suivant les étapes 4 et 5. Le cycle d'incorporation des 4 bases porte le nom de « quad », le processus engendre 25 à 30 quads. D'après (Harris et al., 2008).

La méthode implique la fragmentation de l'échantillon d'ADN, la dénaturation et la polyadénylation en 3', avec à l'extrémité une Adénosine portant une étiquette fluorescente (Fig. 29).

Les brins polyadenylés dénaturés sont hybridés sur des oligonucléotides poly(dT) fixés sur la surface de la cellule d'écoulement ou « flow-cell », avec une densité de 100×10^6 brins d'ADN/cm^2 (Harris et al., 2008). Après enregistrement des coordonnées de positions des brins fixés sur le flow-cell par une caméra CCD grâce a une étiquette fixée à l'extrémité du polyA, celle-ci est clivée et éliminée avant le séquençage. Pour le séquençage, la polymérase et un seul des quatre nucléotides marqué (dNTP-Cy5) est ajouté sur le flow-cell. La caméra CCD permet de déterminer son incorporation dans les brins individuels. Après clivage des étiquettes Cy5 et lavage, le processus est répété avec le dNTP-Cy5 suivant. Chaque cycle de séquençage, qui consiste en l'addition successive de polymérase et des 4 dNTPs étiquetés, est dit « quad ». Le nombre de quads effectué est de l'ordre de 25 à 30, avec des longueurs de lectures allant jusqu'à 45-50 bases. La plateforme Helicos a été utilisée pour séquencer les 6407 bases du génome du bactériophage M13 (Harris et al., 2008).

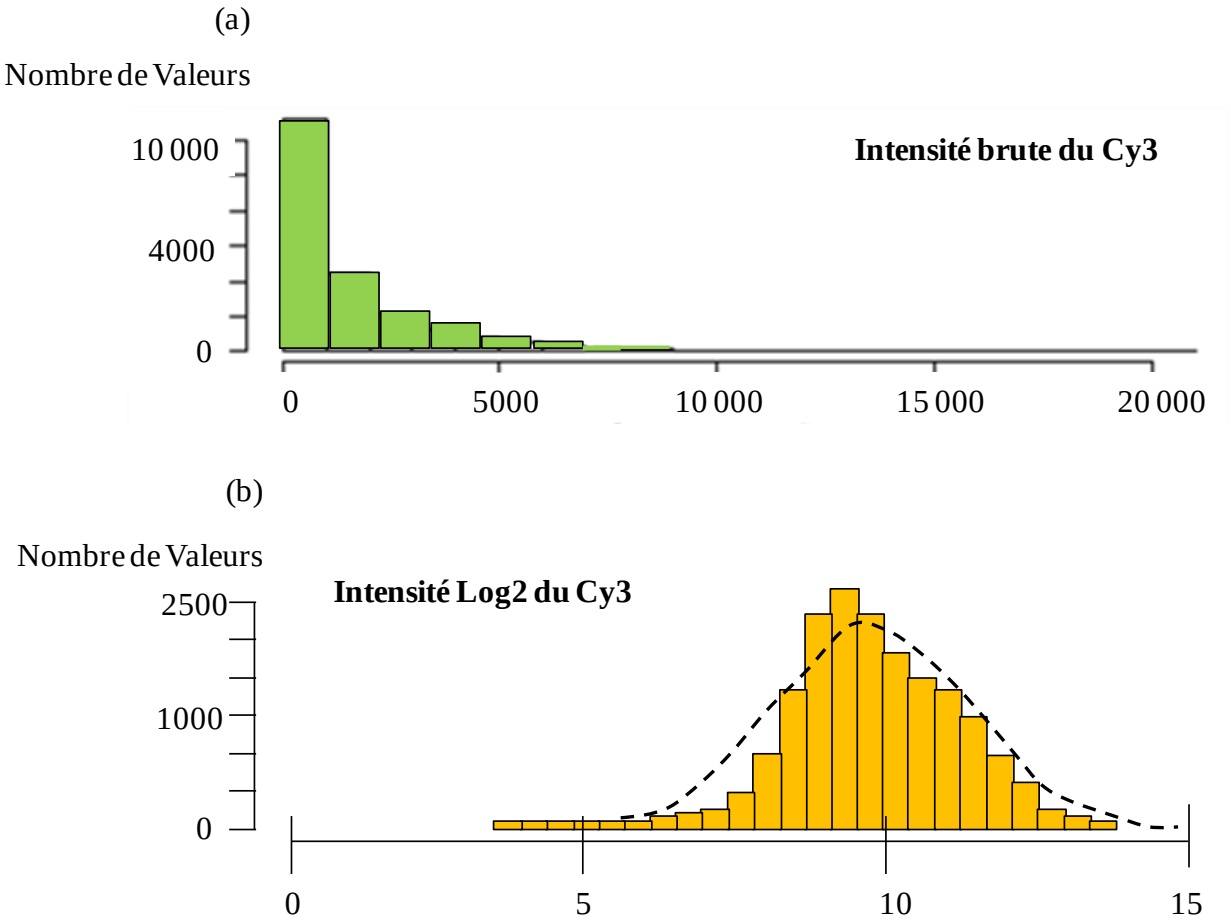


Figure 30: Histogrammes des intensités brutes et log de la Cyanine3.

Données de microarray sur des fibroblastes humains d'après (Stekel, 2003). a) les intensités brutes de la Cy3 sont décalées vers la gauche, avec une majorité de valeurs à de faibles intensités ; et une diminution du nombre de valeurs à de fortes intensités. b) les intensités log de la Cy3 suivent la distribution d'une loi normale (en pointillés). Il existe toujours un faible décalage vers la droite, mais les données log sont plus pertinentes pour l'analyse de données.

1.3.2. Des données haut-débit à l'analyse de l'expression des gènes

Dans cette partie, je présente essentiellement les données issues des deux technologies principalement utilisées dans ma thèse: la technologie Genechip Microarray Affymetrix et le séquençage direct des ARNm (RNA-Seq) utilisant l'outil Illumina 454.

1.3.2.1. Les données et leur pré-traitement

1.3.2.1.1. Des données Genechip Microarray aux profils d'expression de gènes

1.3.2.1.1.1. Analyse d'image

Après hybridation, les fluorochromes sont excitées à la longueur d'onde appropriée et émettent une lumière proportionnelle à la quantité d'ADN cible fixée sur les oligonucléotides de la puce. La puce est scannée à une haute-résolution, aboutissant à une image composée d'un ensemble de spots fluorescents.

L'analyse d'image convertit l'image de chaque spot en valeurs numériques quantifiant l'expression des gènes. Différentes étapes permettent de localiser les différents spots sur la lame, puis de délimiter la zone d'hybridation, l'étape suivante est la mesure du nombre de pixels pour chaque spot. L'intensité globale de fluorescence pour un spot peut ainsi être calculée, en moyennant l'intensité nette mesurée pour chaque pixel (intensité nette = intensité brute - bruit de fond).

Les valeurs d'intensité pour chacun des spots reflétant l'abondance de chaque transcrit sont enregistrées dans un fichier au format « .CEL », constituant les données brutes.

1.3.2.1.1.2. Données brutes

Les données brutes des niveaux d'expression des gènes ont tendance à suivre une distribution asymétrique avec un petit nombre de valeurs élevées. Il est commun de transformer les données brutes en log base 2, afin de travailler sur des données ajustées à une distribution symétrique et proche d'une distribution normale, les modèles gaussiens étant beaucoup plus faciles à manipuler (Fig.30).

1.3.2.1.1.3. Prétraitement des données

Les sources de variabilité sont présentes à chaque étape expérimentale et se confondent avec le signal biologique à étudier. Les données brutes présentent ces nombreux artefacts.
Le prétraitement des données est l'étape initiale de l'analyse, elle doit permettre de minorer les erreurs systématiques et les biais techniques introduits. La réduction de ces variations artefactuelles permet de mettre en évidence les variations biologiques.

1.3.2.1.1.3.1. La correction du bruit de fond

Il existe plusieurs sources d'erreurs dans les données microarray, notamment dans l'hybridation des sondes.

L'hybridation est un processus réversible, dépendant de plusieurs paramètres : la température, la longueur de la sonde et de la cible, la proportion du contenu en Guanine et Cytosine (GC%), les concentrations en sel et en formamide :

- une température élévée (45°C-65°C) facilite l'hybridation tant qu'elle reste inférieure à la température de dénaturation des liaisons double-brins,
- plus les séquences des sondes et des cibles sont grandes, plus leur affinité d'hybridation est grande, due au nombre de liaisons hydrogène,
- la composition en GC est aussi un facteur important pour l'efficacité d'hybridation, puisque les liaisons GC forment 3 liaisons hydrogènes alors que les liaisons AT en présentent uniquement 2,
- la présence de sel augmente l'ionisation des molécules pour une température moins élevée, la présence de formamide minimise les répulsions électrostatiques, ce qui augmente l'efficacité d'hybridation.

La correction de ce bruit de fond permet d'améliorer la fiabilité des résultats. Cependant, elle est efficace lorsque la différence entre les deux dépôts d'un même gène est plus faible en moyenne après la correction.

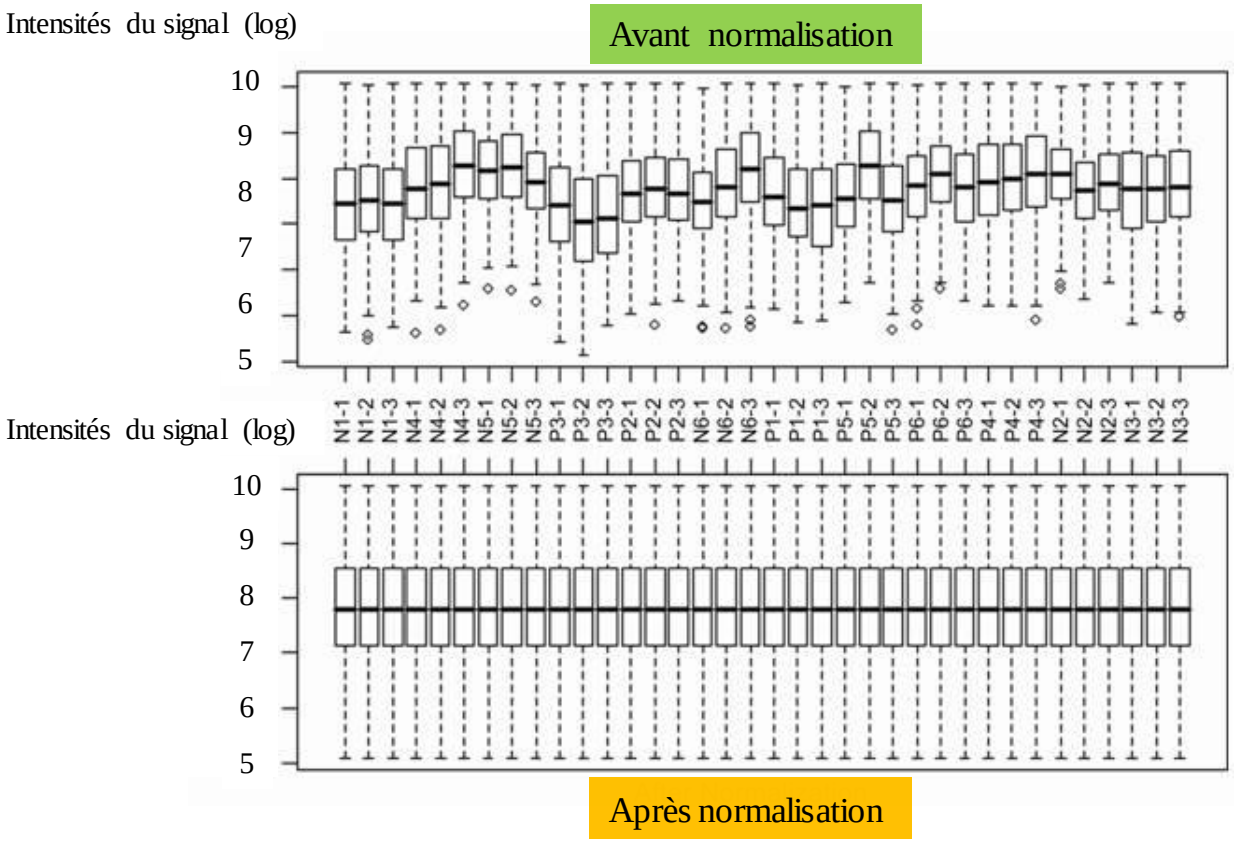


Figure 31: Effet de la normalisation en quantile illustré par Box plots sur un jeu de données.

Chaque box plot correspond à un échantillon (Su et al., 2008).

1.3.2.1.1.3.2. La normalisation

Les biais affectent tous les gènes de façon similaire. La normalisation permet de minorer ces biais en donnant la même moyenne et la même variance à toutes les conditions expérimentales afin de rendre les puces comparables entre elles. Elle permet de donner le même poids aux différentes conditions étudiées dans les analyses qui suivent (Fig. 31). C'est aussi faire l'hypothèse que la quantité totale d'ARNm dans les cellules est constante entre les différentes conditions étudiées.

Plusieurs méthodes de prétraitement des données affymetrix sont décrites dans la littérature. Dans mes travaux de thèse, j'utilise la méthode gcRMA (Guanine Cytosine Robust Multiarray Average) dans l'article publié (Chelaifa et al., 2013), et la méthode RMA (Robust MultiArray Analysis) (correction du bruit de fond et normalisation) dans le chapitre 3.

Les puces Affymetrix Genechips sont conçues avec deux types d'oligonucléotides, les PM (perfect match) et les MM (mismatch) (détaillé précédemment dans le paragraphe 1.3.1.2.1.2.), s'hybridant de manière spécifique (PM) ou non spécifique (MM) à la cible. La soustraction du signal des sondes MM aux sondes PM aboutit à une valeur correcte du signal et permet d'éviter une sur-représentation des produits du gène.
Parfois, le signal des sondes MM est supérieur à celui des sondes PM. Dans ce cas (où signal MM > signal PM), la sonde MM détecte le bruit de fond. Le résultat est alors une valeur d'expression négative (PM – MM < 0, car MM> PM) et l'interprétation est que le véritable produit du gène n'est pas exprimé dans cette condition.

La méthode RMA, pour Robust MultiArray Analysis, est une procédure de normalisation des puces à oligonucléotides, qui corrige le bruit de fond, normalise et récapitule l'information au niveau de la sonde, sans utiliser l'information obtenue des sondes MM. La méthode GC-RMA est une amélioration de la méthode RMA, qui utilise la séquence de chacune des sondes PM et MM pour préciser les valeurs d'expression des gènes.

Correction du bruit de fond RMA et GC-RMA

Dans la correction du bruit de fond RMA (Irizarry et al., 2003), chaque puce est corrigée, indépendamment, au niveau des sondes. Seule la valeur PM sur les sondes est considérée et modélisée, comme la somme du signal d'intérêt et du signal de bruit de fond (dû au bruit optique et aux liaisons non spécifiques): PM = Signal + Bruit de fond.

La méthode GC-RMA (Wu et al., 2004), utilise un modèle qui prend en compte la composition en G/C de chacune des sondes (PM et MM) à la surface de la puce. Sous l'hypothèse d'une affinité d'hybridation plus forte entre les sondes MM et les cibles quand la composition en G/C est grande, le bruit de fond est donc proportionnel à la composition en G/C dans chacune des sondes.
Cette correction du bruit de fond prend également en compte la position de chacune des bases (A, T, C ou G) le long de la sonde et détermine l'affinité de chaque sonde.

Après la correction du bruit de fond, l'étape de la normalisation permet de comparer et d'optimiser les niveaux d'intensité entre les différentes conditions.
Les procédures RMA et GC-RMA comportent une normalisation en quantile (Bolstad et al., 2003), c'est-à-dire que les signaux sont ajustés de manière à ce que toutes les puces aient la même distribution d'intensité des sondes, car on considère que la majorité des gènes ne sont pas différentiellement exprimés entre plusieurs conditions. Pour chacune des puces, les données corrigées sont normalisées par rapport aux quantiles dérivés de l'ensemble des données.

La normalisation GC-RMA est la méthode la plus célèbre pour convertir les données brutes microarray en profils d'expression de gènes, et semble être l'une des meilleures procédures de normalisation pour la détection de gènes différentiellement exprimés (Wu et al., 2004).

Après l'ajustement du bruit de fond et la normalisation, les données issues d'un jeu de sondes pour la même cible (probeset) sont résumées. Les intensités des sondes d'un probeset donné sont combinées pour définir une valeur d'expression par probeset.

1.3.2.1.2. Des données Illumina à l'expression des gènes

Au cours de mes travaux de thèse, j'ai aussi utilisé des données de séquençage direct des ARNm sur l'espèce complexe allopolyploïde blé, ce qui sous entend la mise au point de méthodes d'analyses.

1.3.2.1.2.1. Traitement d'image

Le processus ou « run » de séquençage génère des séries d'images analysées par le logiciel Illumina. Chaque cluster, caractérisé par une position (X ; Y), est d'abord repéré dans chacun des fichiers d'image. Puis le signal d'intensité et le bruit de fond sont mesurés. Les intensités brutes sont converties, par une identification des bases ou « base calling », en séquences nucléotidiques (pour une longueur comprise entre 35 et 101 pb) appelé lecture ou « read ». Cette lecture au format fasta est associée à un score de qualité (« phred score » codé en ASCII), qui permet de caractériser la qualité de la séquence. Les scores de qualité varient entre 4 et 60, les valeurs élevées correspondent à une meilleure qualité. Le score « phred »est logarithmiquement lié aux probabilités d'erreurs, tel que:

$$Q = -10 \log_{10} P$$

soit :

$$P = 10^{-Q/10}$$

avec P, la probabilité d'erreur, et Q le score de qualité phred.

Ainsi, un score de qualité 10, correspond à une probabilité d'erreur $P=10^{-1}$, soit à 1 erreur pour 10 bases.

Ces informations (de séquences et de scores de qualité) sur la lecture sont associées dans un fichier au format « fastq ».

1.3.2.1.2.2. Alignement des lectures et résumé des lectures alignées

Afin de quantifier le niveau d'expression des transcrits, les lectures de données NGS sont alignées par similarité sur un génome de référence où les séquences sont annotées dans la totalité (ou au minimum pour les gènes). C'est une des étapes fondamentales pour l'analyse des données (Bao et al., 2011). Le nombre de lecture par gène est par la suite comptabilisé pour apprécier la quantité exprimé.

De nombreux outils logiciels ont été développés pour l'alignement et le comptage de ces lectures sur des séquences de référence, suscitant des comparaisons de performances en termes de sensibilité, précision, vitesse et exigence en mémoire vive (Bao et al., 2011).

Les lectures de données NGS (typiquement 35–400 bp comparées aux lectures basées sur la technologie de Sanger de 650–800 bp) présentent deux caractéristiques principales:

- un nombre important de données, nécessitant un grand espace de stockage et une grande vitesse de traitement, donc une utilisation optimale de la mémoire et de la vitesse de calcul des serveurs ;

- la présence de multiples profils d'erreurs dans les données liées aux différentes technologies de séquençage.

Les méthodes classiques d'alignement (BLAST[12] et BLAT[13]) aligneraient les lectures en quelques jours (grâce à un serveur de calcul informatique performant). Ces méthodes classiques reposent sur l'algorithme d'alignement de séquence de Smith-Waterman (inventé en 1981), qui donne un alignement correspondant au meilleur score possible entre deux séquences à comparer (protéiques ou nucléiques). Par rapport à l'algorithme de Needleman-Wunsch (autre algorithme d'alignement de séquences), l'algorithme de Smith-Waterman offre l'avantage de rechercher, en plus des alignements globaux[14], des alignements locaux[15], n'impliquant que des régions ou des segments entre deux séquences analysées.

[12] BLAST (Basic Local Alignment Search Tool) est une méthode heuristique de comparaison de séquences, fondée sur la méthode de Smith-Waterman (alignement local).

[13] BLAT (BLAST-Like Alignment Tool) est une version très rapide de BLAST.

[14] alignement des séquences sur toute leur longueur.

[15] alignement des séquences sur une partie de leur longueur

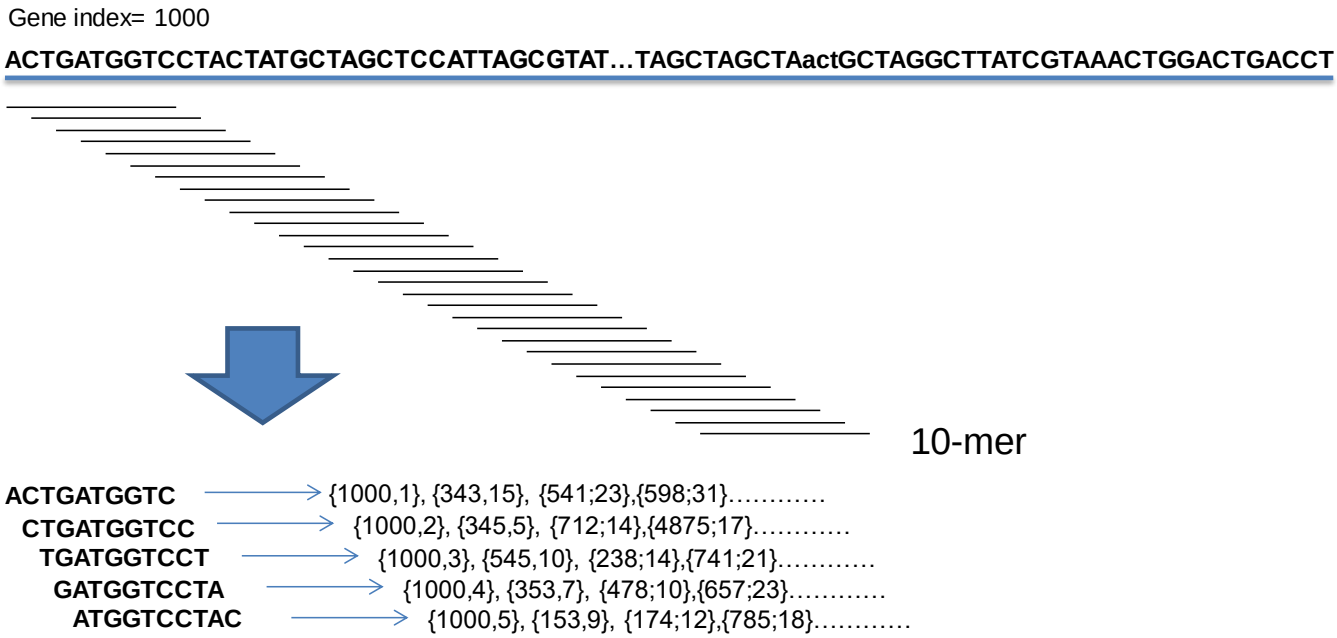


Figure 32: Schéma d'une table de hachage.

La table de hachage est une structure de données qui contient l'ensemble des oligonucléotides de 10 pb (10-mer) dans l'ensemble de données. La clé de la table est une séquence de 10pb, la valeur de hachage est une liste de couple de valeurs indiquant l'indice de la séquence et les positions de l'oligonucléotide de 10pb sur cette séquence. La table de hachage utilise une fonction qui prend en entrée la séquence de l'oligonucléotide de 10 pb et renvoit la valeur de hachage.

L'utilisation exponentielle des technologies NGS a suscité le développement de nombreuses méthodes d'alignements spécifiques aux données produites (Bozdag et al., 2009; Chen et al., 2009a; Schatz, 2009; Clement et al., 2010) : SOAP2 (Li et al., 2009), Bowtie(Langmead et al., 2009), Bowtie2 (Langmead and Salzberg, 2012), BWA(Li and Durbin, 2009), MAQ(Li et al., 2008a), ZOOM(Lin et al., 2008), SHRiMP(Rumble et al., 2009), PerM (Chen et al., 2009b), RMAP(Smith et al., 2009), SeqMap(Jiang and Wong, 2008), BFAST(Homer et al., 2009), MOM (Eaves and Gao, 2009), MosaikAligner (Lee et al., 2014), NovoAlign (http://www.novocraft.com), PASS (Campagna et al., 2009), ProbeMatch (Kim et al., 2009), SSAHA2 (Ning et al., 2001), GSNAP(Wu and Nacu, 2010), FANGS (Misra et al., 2009), mrFAST (Alkan et al., 2009), mrsFAST (Hach et al., 2010), SOAP3 (Liu et al., 2012a)…

Ces outils d'alignements de séquences peuvent être classés selon leurs méthodes d'indexation faisant appel à deux types d'algorithmes:

- les algorithmes fondés sur une table de hachage,
- les algorithmes fondés sur la transformée de Burrows-Wheeler (BWT).

Une table de hachage est une structure de données suivant une indexation clef-élément (Fig.32). L'accès à un élément se fait à partir de sa clef, qui est transformée en une valeur de hachage (ou index de l'élément) par une fonction.
Les méthodes sur le hachage sont divisées en deux types:

- le hachage des lectures,
- le hachage du génome.

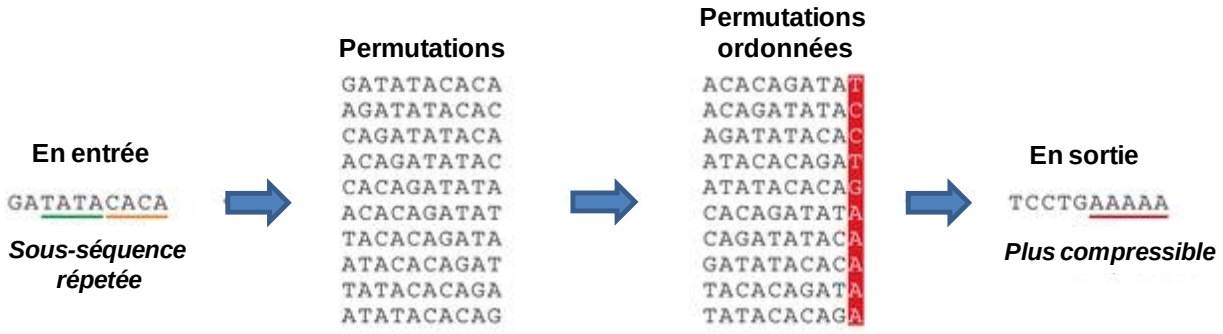


Figure 33: Technique d'indexation pour le séquençage de données par BWT.

D'après (Berger et al., 2013), la transformée de Burrows–Wheeler est la transformation de chaîne de caractères. Les sous-séquences fortement repétées sont converties dans un format qui peut-être facilement compressé. Ceci est réalisé par : 1)la génération de toutes les permutations possibles de la sous-séquence donnée en entrée, tel que chacune des positions devient la position de début une seule fois ; 2) les séquences permutées sont alors ordonnées dans l'ordre alphabétique et 3) la dernière colonne de toutes les permutations est extraite et correspond à la séquence en sortie. Si il ya des sous-séquences répétées dans la séquence en entrée, la transformation BWT regroupe ces sous-séquences et aboutit à la répétition d'un seul caractère dans la dernière colonne ; par conséquent, la partie répétée de la séquence en sortie facilite la compression.

Pour ces deux méthodes, l'idée générale est de construire une table pour les sous-séquences des lectures ou du génome. La clef de chaque entrée est la sous-séquence et l'élément de la table est la liste des positions de la sous-séquence.
Parmis les programmes cités précédemment on trouve l'implèmentation d'une table de hachage des lectures pour: Eland, Maq, RMAP, SeqMap, mrFAST /mrsFAST, SHRiMP, ZOOM, CloudBurst, GSNAP (Bao et al., 2011).
Tandis que les programmes : MOM, MOSAIK, PASS, ProbeMatch, SOAP, SSAHA2, Novoalign, PerM, BFAST, FANGS implémentent une table de hachages du génome, qui est stockée pour les comparaisons avec les lectures (Bao et al., 2011).

La transformation de Burrows-Wheeler (noté BWT)(Burrows and Wheeler, 1994) repose sur une réorganisation des données pour atteindre un meilleur taux de compression (Fig.33). Les données réorganisées sont appelées « transformée de Burrows-Wheeler ». Plusieurs outils d'alignement de lectures implémentent cette transformation afin d'utiliser moins de mémoire lors de l'alignement des lectures sur le génome de référence (Martin and Wang, 2011).
BWT a été améliorée par Ferragina et Manzini (Ferragina and Manzin, 2000) en FM-index. La transformation du génome par la méthode FM-index permet de gagner en performance dans les cas où une même lecture s'aligne à plusieurs endroits sur le génome de référence.
Parmis les outils cités précédement : BWT, SOAP2, Bowtie, Bowtie2, BWA, SOAP3 utilisent la méthode de BWT.

Les outils d'alignement basés sur BWT sont plus sensibles mais restent moins rapides que ceux utilisant une table de hachage pour l'alignement des séquences (Ferragina and Manzini, 2005; Lee et al., 2014). Mais, les outils fondés sur la table de hachage consomment généralement plus de mémoire que ceux utilisant la méthode de BWT (Lee et al., 2014).

Au cours de mes travaux de thèse, j'ai essentiellement utilisé deux logiciels d'alignements de lectures, fondés sur la méthode d'alignement de BWT : SOAP2 et BWA.

Le logiciel **SOAP2** utilise la transformée de Burrows-Wheeler bidirectionnelle (2way-BWT), pour indexer le génome de référence et accélérer le processus d'appariement exact ou « exact matching » (ou identité). La transformée de Burrows-Wheeler bidirectionnelle permet de repérer le motif dans les 2 sens (recherche en amont et aval de la séquence alignée).
La principale caractéristique du logiciel SOAP2 est sa performance dans la détection de SNP, ce qui explique son utilisation pour ce type de recherche. Son point faible selon la littérature (Schbath et al., 2012), est qu'il remplace, dans les lectures, tous les 'N' par un « G », ce qui peut-être la cause d'erreurs lors de l'analyse.

Aujourd'hui, le logiciel BWA est très utilisé pour l'alignement de génomes eucaryotes, car il présente de nombreux avantages : notamment une plus grande souplesse (alignement des lectures avec ou sans gap –ou interruption de séquences) et un temps de calcul plus court.
BWA utilise la transformée de Burrows-Wheeler, qui permet de repérer un motif dans un seul sens (recherche en amont).

Le logiciel BWA compte trois algorithmes : BWA-backtrack, BWA-SW et BWA-MEM (http://bio-bwa.sourceforge.net/bwa.shtml). BWA-backtrack est conçu pour les lectures de petites séquences, type Illumina (jusqu'à 100pb), BWA-SW et BWA-MEM sont conçus pour les séquences plus longues (entre 70pb et 1000pb). BWA-MEM est le plus récent, le plus rapide et le plus précis, il utilise un algorithme plus performant (Li, 2013)(pour les lectures Illumina entre 70-100pb). Dans mes travaux de thèse, j'utilise la chaîne de traitement informatique interne du Genoscope qui utilise BWA-backtrack.
Il est à noté que BWA remplace dans le génome de référence les 'N' par un nucléotide (aléatoirement A, C, T ou G).
Ces deux logiciels d'alignements de lectures (SOAP2 et BWA) que j'ai utilisé sont réputés pour être les plus rapides actuellement (Shang et al., 2014).

Pour conclure sur les logiciels d'alignement pour les données issus des méthodes NGS : BWA et SOAP2 (ou Bowtie, Mosaik, SHRiMP) présentent des résultats d'alignements satisfaisants pour les lectures Illumina paired-end (PE) ou single-end (SE), toutefois BWA utilise le moins de mémoire du fait de son algorithme BWT et SOAP2 est le plus rapide en calcul (du fait de son algorithme (2way-BWT)) (Bao et al., 2011 ; Lindner and Friedel, 2012 ; Shang et al., 2014).

Ces méthodes sont différentes sur la sensibilité d'alignement, qui pourrait être liée à l'heuristique[16] utilisée pour la détection de mismatch (Horner et al., 2009 ; Bao et al., 2011). BWA est plus fréquemment utilisé pour repérer les SNP et les indel (insertions/deletions), puisqu'il introduit des gaps lors de l'alignement avec peu d'erreur.

En général, pour l'alignement exact des lectures la méthode de BWT reste plus rapide que les tables de hachages, car elle compresse les séquences du génome et écrase les régions répétées en une seule copie (Schbath et al., 2012). Une séquence est alors alignée à toutes les copies plutôt qu'à chacune des copies tel que dans une table de hachage (biostars, 2011). Pour les alignements avec mismatch la méthode de BWT devient moins performante, notamment pour les lectures de 100pb où les tables de hachage sont plus rapides (biostars, 2011).

Le logiciel SOAP2 supporte de multiples formats de fichiers en entrée (fasta ou fastq) et génère en sortie un fichier au format SAM (Sequence Alignment/Map), qui est un format d'alignement pour le stockage d'alignements de lectures sur des séquences de référence (Service, 2006 ; Li et al., 2009). Le logiciel BWA prend également en entrée un fichier fasta ou fastq et renvoie en sortie un fichier au format SAM. Le format SAM nécessite l'utilisation du logiciel samtool pour son traitement.

Le format BAM est une version compressée (binaire) du format SAM ce qui la rend plus rapide pour l'accès aux données. Ces formats supportent les lectures courtes et longues produites par différentes plateformes de séquençage, et permettent de stocker toutes les informations d'alignement.

[16] méthode de calcul qui fournit rapidement une solution réalisable, pas nécessairement optimale ou exact

Une fois les lectures alignées sur les séquences référence, celles-ci sont récapitulées dans un fichier de sortie, après le déroulement de plusieurs scripts aboutissant à un tableau décrivant pour chacune des séquences référence le nombre de comptages brut obtenu pour chacune des conditions testées.

1.3.2.1.2.3. Normalisation

La comptabilisation des données RNA-seq présente des biais intra-échantillon et des biais inter-échantillons.

Parmi les biais intra-échantillon, on constate que le nombre de lectures alignées augmente avec la longueur du gène (Oshlack and Wakefield, 2009) et que la composition de la séquence en GC influe aussi sur le comptage des lectures (Benjamini and Speed, 2011 ; Risso et al., 2011; Zheng et al., 2011). En effet, les séquences riches ou pauvres en GC sont sous-représentées dans les RNA-Seq car très peu de lectures s'alignent sur ces 2 types de régions. Ainsi, dans un même couloir (ou lane en anglais) du flow-cell les comptages de lectures ne sont pas directement comparables entre les gènes. On constate également que les gènes hautement exprimés dans un échantillon représentent une plus grande proportion des lectures séquencées, ce qui a pour conséquence de réduire le nombre de lectures alignées sur les autres gènes.

Parmis les biais inter-échantillons : la taille de la librairie influe sur le nombre de comptages, car le nombre de lectures alignées augmente avec la taille de la librairie (Tarazona et al., 2011), et peut donc varier d'un échantillon à l'autre.

Pour corriger ces biais, il existe deux types de méthodes :

- la normalisation inter-librairie (pour chaque échantillon) corrige les biais inter-échantillon pour comparer les gènes entre les différents échantillons,
- la normalisation intra-librairie (pour chaque gène) corrige les biais intra-échantillon permettant la comparaison des gènes d'un même échantillon.

Je présente ici la normalisation RPKM, que j'utilise dans mes travaux et son analogue FPKM.

RPKM

La normalisation RPKM (« Reads Per Kilobase of exon per Million mapped reads » traduit par « lectures par kilobase d'exon par million de lectures alignées ») (Mortazavi et al., 2008) fait l'hypothèse que le nombre de lectures alignées sur un gène est proportionnel :

- à la longueur du gène,
- à son niveau d'expression,
- à la taille de la librairie d'où il provient.

Les données normalisées sont obtenues après division des données de comptage par un facteur de normalisation, noté F_{ij} égal au produit du nombre total de lectures alignées dans l'échantillon j (en millions de lectures) par la longueur du gène i (en kilobase) :

$$F_{ij} = \frac{N_j}{10^6} \times \frac{L_i}{10^3}$$

avec :

N_j le nombre de lectures alignées dans l'échantillon j,

L_i la longueur du gène (ou des exons du gène) i.

La normalisation **FPKM** ou « Fragments Per Kilobase of transcript per Million mapped reads » (fragments par kilobase de transcrits par million de fragments alignés) est analogue des RPKM pour les données pairées (Trapnell et al., 2010).

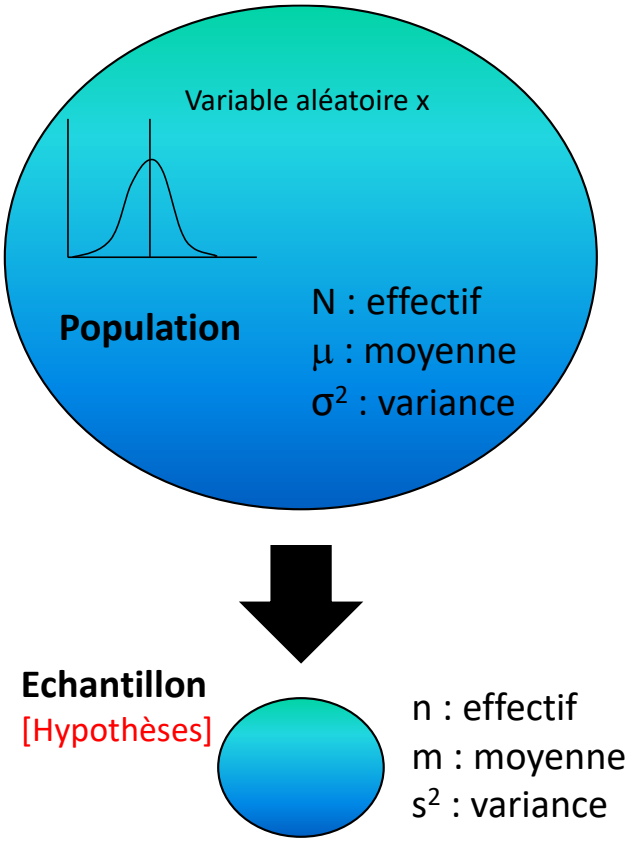


Figure 34: Schéma d'un échantillon issu d'une population.

Dans la population dont est issu l'échantillon, on pose l'hypothèse que la variable aléatoire x, suit une loi Normale de moyenne μ et de variance σ^2. Dans l'échantillon de taille n, x a une moyenne m et une variance s^2.

1.3.2.2. Analyse statistiques des données transcriptome

Au cours de mes travaux de thèse, l'analyse de données transcriptomiques commence par le pré-traitement des données (détaillé précédemment par l'étape de normalisation des données), puis se poursuit par une analyse différentielle.

Dans cette partie, je rappelle quelques notions élémentaires de statistiques nécessaires à la compréhension de la méthode d'analyse différentielle utilisée au cours de mes travaux de thèse.

1.3.2.2.1. Les tests d'hypothèses

Considérons une population présentant un caractère (quantitatif ou qualitatif) à étudier, dont la valeur du paramètre est inconnue. On formule une hypothèse sur la valeur de ce caractère d'intérêt, et on porte un jugement sur cette hypothèse, sur la base des résultats d'un échantillon prélevé de cette population (Fig.34).

1.3.2.2.1. Concepts utiles aux tests d'hypothèse

Hypothèse statistique

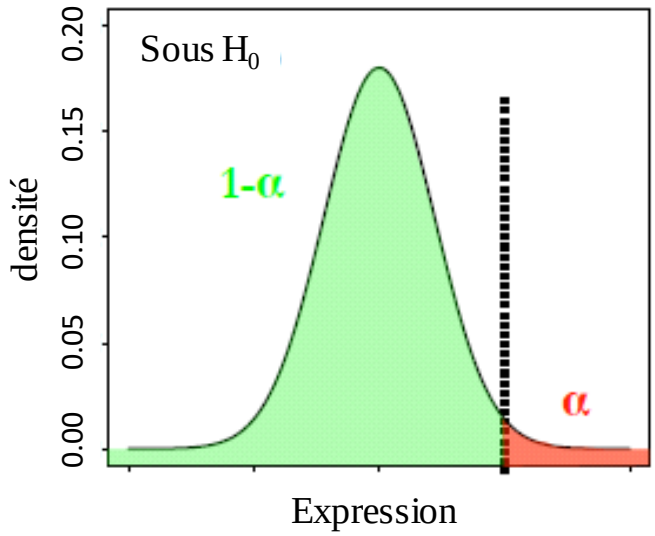
Une hypothèse statistique est une affirmation sur le modèle statistique utilisé pour décrire les caractéristiques d'un échantillon (valeurs des paramètres, distribution des observations).

Test d'hypothèse

Un test d'hypothèse (ou test statistique) est une démarche consistant à fournir une règle de décision, en fonction d'une réalisation observée de l'échantillon. Ce test permet de décider quelle l'hypothèse est soutenue par les observations.

Hypothèse nulle et hypothèse alternative

L'hypothèse nulle, notée H_0, fixe *a priori* le paramètre de la population à une valeur donnée. L'hypothèse alternative, notée H_1, est l'hypothèse différente de l'hypothèse H_0.



.

Figure 35: Schéma du seuil de signification de test α.

D'après (Aubry).

Réalité/Vérité

Décision	H_0 vraie (H_1 fausse)	H_0 fausse (H_1 vraie)
H_0 acceptée (H_1 rejetée)	Bonne décision (1-α) **Vrais Positifs**	Erreur β **Faux Négatifs** • risque de 2ème espèce(risque β); • risque de se tromper quand on **accepte H_0**
H_0 rejetée (H_1 acceptée)	Erreur α **Faux Positifs** • risque de 1ère espèce (risque α); • risque de se tromper quand on **rejete H_0**	Bonne décision (1-β) **Vrais Négatifs** • (puissance du test) • capacité du test à identifier une différence réelle

Tableau 1 : Réprésentation des erreurs α et β, des vrais positifs et des vrais négatifs.

Statistique de test

La décision d'accepter ou rejeter l'hypothèse H_0 repose sur une statistique de test.

La statistique de test est une variable d'échantillonnage, qui permet d'estimer la valeur du paramètre dans la population. Cette statistique de test n'est pas strictement égale à la valeur théorique énoncée dans le test d'hypothèse, due aux fluctuations d'échantillonnage. Pour décider si l'hypothèse formulée correspond aux observations, il faut vérifier si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre dans l'hypothèse est trop grand pour être due uniquement au hasard.

Seuil de signification du test

Dans la démarche de test statistique, la règle de décision conduit à rejeter ou non l'hypothèse H_0. Le test étant réalisé sur un échantillon de la population, la décision comprend un risque appelé « le seuil de signification » ou « niveau de significativité », noté α.

Le risque α ou erreur de première espèce est le risque de se tromper en rejetant H_0, alors que H_0 est vraie. Le seuil de signification s'énonce : $\alpha = P(\text{rejeter } H_0 \mid H_0 \text{ vraie})$ (Tableau 1).

Sur la distribution d'échantillonnage de la statistique de test, ce seuil de signification correspond à une région de rejet de l'hypothèse H_0 (Fig. 35). Sur cette même distribution, une région complémentaire de probabilité 1-α correspond à la région de non-rejet de H_0. Les seuils de signification les plus courants sont α=1%, 5% ou 10% selon l'étude.

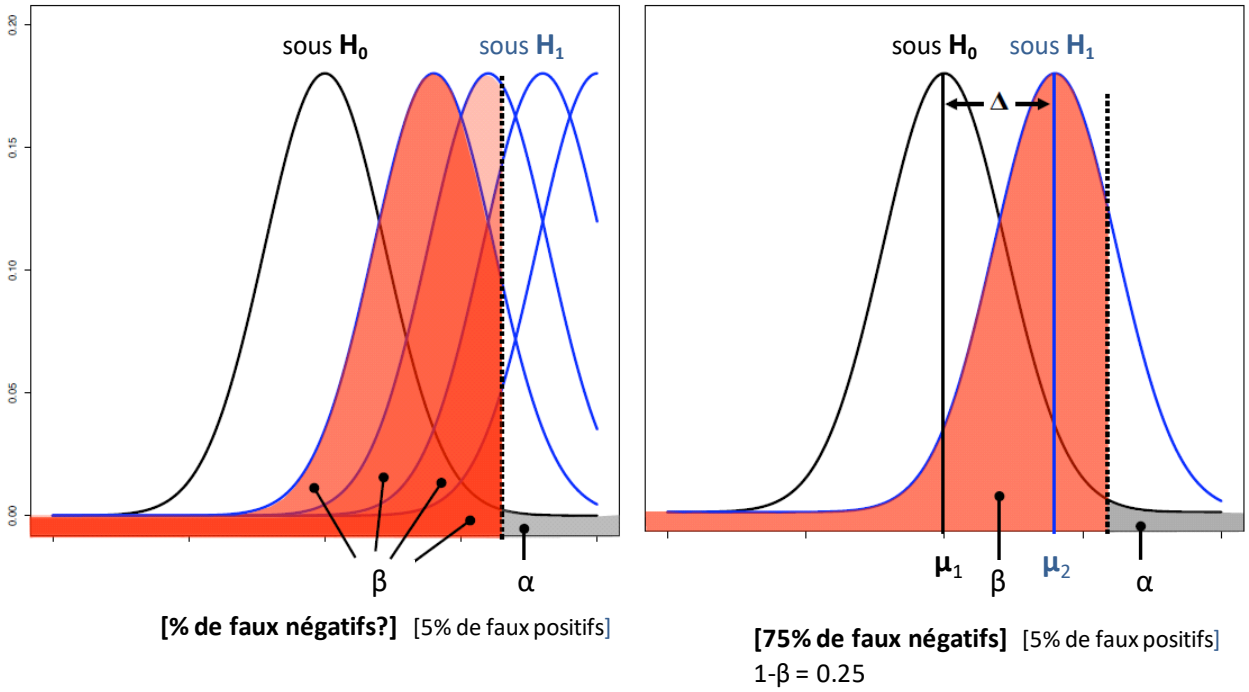


Figure 36: Schéma des erreurs α et β sous la distribution de H_0.

D'après (Aubry).

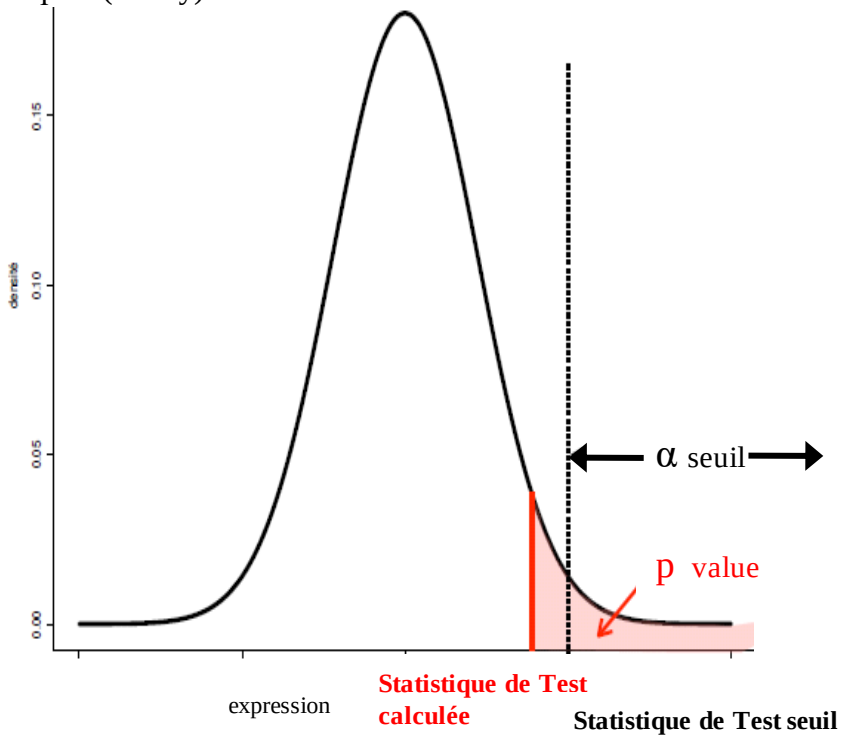


Figure 37: Représentation de la p-value.

Erreurs

Lors de la conclusion d'un test d'hypothèse, « accepter H_0 alors que H_0 est vraie » est une bonne décision, « rejeter H_0 alors que H_0 est fausse » est aussi une bonne décision (Tableau 1). Toutefois, il existe deux types d'erreurs :

- l'erreur α (comme décrit précédemment) ou erreur de type I est le risque de rejeter H_0 alors que H_0 est vraie. Cette erreur de type I correspond aux faux Positifs,
- l'erreur β ou erreur de type II (ou erreur de deuxième espèce) est le risque de se tromper en acceptant H_0 alors que H_0 est fausse. Cette erreur correspond aux faux Négatifs.

Ces deux erreurs α et β ne sont pas indépendantes, mais inversement liées : lorsque l'une augmente l'autre diminue et inversement (Fig. 36). On choisit de contrôler en priorité le risque α, telle que la probabilité maximale de faire une erreur de type I reste inférieure ou égale à α.

Puissance de test

La puissance de test est une variable égale à 1-β (Tableau 1), elle correspond à la capacité de rejeter H_0 quand H_0 est fausse. En d'autres termes, c'est l'aptitude à mettre en évidence une différence lorsqu'elle existe. La puissance de test dépend (Fig.36):

- du seuil α choisi, la puissance diminue quand le seuil α décroit,
- de la variance de la population σ^2 la puissance diminue quand σ^2 augmente,
- de l'effectif (n) de l'échantillon, la puissance augmente quand n augmente,
- de l'écart (Δ) entre les paramètres testés, la puissance augmente quand Δ augmente.

p-valeur

La p-valeur (ou « p-value ») est le risque que l'on prend à rejeter H_0. La p-valeur correspond à l'aire sous la courbe de la distribution de la statistique de test sous H_0 (Fig. 37).

1.3.2.2.2. Analyse différentielle de données transcriptomique

L'analyse différentielle consiste en l'identification de gènes différentiellement exprimés entre plusieurs conditions expérimentales, afin de sélectionner des gènes spécifiques d'une condition. Toute analyse différentielle repose sur des tests statistiques, le principe est le même pour les données microarray ou RNA-Seq.

La première étape commence par la définition des hypothèses H_0 et H_1, où H_0 pose l'hypothèse de l'égalité entre les conditions, et H_1 l'hypothèse alternative.
Puis, on définit la valeur du seuil de signification du test α. Le seuil de signification le plus couramment utilisé est α =5%, c'est aussi le seuil utilisé tout au long de mes travaux de thèse.
L'étape suivante est de définir le bon test statistique. Parmi ces tests, le t-test est le plus connu, il consiste à comparer des moyennes. Deux t-test différents sont appliqués dans l'analyse des données microarrays (présenté dans le chapitre 3) :

- le test de Student considère l'homogénéité des variances de deux échantillons à comparer et utilise une variance pondérée moyenne de ces deux variances,
- le test de Welch considère l'hétérogénéité des variances de 2 échantillons à comparer et utilise les variances de ces deux échantillons.

La statistique de test et la p-valeur associée sont alors calculées. En pratique, on utilise la p-valeur pour décider de rejeter ou non H_0. Une p-valeur inférieure au seuil de signification α entraine le rejet de H_0, une p-valeur supérieure ou égale entraine le non-rejet de l'hypothèse H_0. On dit plutôt « ne pas rejeter H_0 » plutôt qu' « accepter », car une p-valeur supérieure ou égale au seuil α ne signifie pas l'absence d'une différence ou une égalité des moyennes dans les populations dont sont extraits les échantillons.

Soit X_i une variable aléatoire valant 1 si le gène n'est pas un faux positif, 0 sinon. X_i suit une loi de Bernoulli de paramètre 5%. Le nombre total de faux positif parmi 100 est une variable aléatoire binomiale S_n égale à :

$$S_n = \sum_{i=1}^{100} X_i \sim \mathrm{b}(100, 5)$$

dont la loi est donnée par :

$$\mathbb{P}(S_n = k) = C_n^k 0.05^k 0.95^{n-k}$$

La probabilité que le nombre total de faux positif soit supérieur ou égale à 1:

$$\mathbb{P}(S_n \geq 1) = 1 - \mathbb{P}(S_n = 0) = 1 - 0.95^{100} \approx 0.994 = 99.4\%.$$

1.3.2.2.3. Tests multiples

Les données à haut-débit, telles que les microarrays, ou les RNA-Seq, concernent plusieurs dizaines de milliers voire des millions de gènes simultanément. Ainsi, lorsqu'un test statistique est appliqué à chacun de ces gènes et que l'on considère les résultats des tests dans leur ensemble, un problème de comparaison multiple – ou de « tests mutiples » se pose (voir par exemple (Jeanmougin, 2012) dans le contexte de la génomique).

Imaginons en effet qu'un gène soit déclaré différentiellement exprimé entre deux conditions, c'est-à-dire que l'on rejette l'hypothèse nulle au niveau α=5%. Si l'on s'intéresse uniquement à ce gène, il n'y a que 5% de chance d'avoir rejeté l'hypothèse nulle (le gène est non différentiellement exprimé) à tort. On contrôle donc le taux de faux positifs à 5%. En revanche, si l'on s'intéresse au sort de 10 000 gènes et que l'on déclare 100 de ces gènes comme différentiellement exprimés, le nombre de faux positifs est de 5% x 100 = 5 et la probabilité d'avoir au moins un faux positif de 99.4 % (sous l'hypothèse d'indépendance entre les tests) (détaillé ci-contre).

Différentes techniques ont été développées afin de contrôler le taux de faux positifs associé à un ensemble de tests – ou une grandeur associée -, par exemple en utilisant le critère FWER (Family-wise error rate) et/ou FDR (False Discovery error Rate). On parle de « correction de tests multiples » lorsque l'on modifie les probabilités critiques (p-valeurs) des tests afin de contrôler un de ces critères, qui porte sur la totalité des tests.

Le contrôle du FWER :

Le FWER (ou Family-Wise Error Rate) est la probabilité d'avoir au moins une erreur de type I, ou encore, dans ce contexte de recherche, la probabilité de sélectionner à tort un ou plusieurs gènes sur l'ensemble des gènes étudiés (Hochberg and Tamhane, 1987). Le FWER contrôle la probabilité de n'avoir aucun faux positif. Mécaniquement, les procédures contrôlant ce critère ont tendance à être peu puissantes, c'est-à-dire à présenter un taux élevé de faux négatifs. Ainsi, les procédures contrôlant ce critère sont d'autant moins puissantes que le nombre d'hypothèses testées est important (méthodes : Bonferroni, Sidak, Holm).

Bonferroni

Selon la correction de Bonferroni, si les k tests sont effectués avec un seuil de rejet de H_0 à α_{seuil}/k, la probabilité de faire une erreur sur l'ensemble des k tests est au maximum de α_{seuil}. Si la limite de rejet sur quatre tests est de 0.05/4=0.0125, le risque pour les quatre tests est de rejeter H_0 alors que la probabilité que H_0 soit vrai est de 5%.

Contrôle des FDR :

Benjamini-Hochberg, en 1995, proposent le critère d'erreur correspondant au "taux de fausses découvertes" ou « False Discovery Error Rate » (FDR) (Benjamini and Hochberg, 1995). Le FDR contrôle l'espérance (la moyenne) de faux positifs dans les gènes différentiellement exprimés. Le FDR détecte plus de vrais positifs que le FWER, mais contrôle moins strictement le taux de faux positifs. Il est plus adapté pour les études de génomique fonctionnelle et s'est imposé comme le critère standard dans ce contexte.
Il existe plusieurs méthodes pour le contrôle du FDR dont : Benjamini-Hochberg, Benjamini-Yektutieli que j'utilise dans mes travaux de thèse.

Benjamini-Hochberg

La procédure de Benjamini-Hochberg (BH, (Benjamini and Hochberg, 1995)) repose sur le classement en rangs des différentes p-valeurs avant de les ajuster. Elle contrôle le FDR au niveau α pour m tests indépendemment effectués.
La procédure BH tient en trois étapes :

i) ranger les p-values dans l'ordre croissant et les numéroter (de 1 à m);

ii) pour un niveau α donné, trouver le plus grand j, tel que :

$$p_j \leq \frac{j}{m}\alpha$$

iii) rejeter les H_0, pour les tests allant de 1 à ce j.

Benjamini-Yektutieli

La procédure de Benjamini-Hochberg-Yekutieli (Benjamini and Yekutieli, 2001) est une modification de Benjamini-Hochberg sous l'hypothèse de tests non indépendants. Un facteur c(m) est ajouté au critère défini dans l'étape ii) de la procédure BH, tel que :

$$p_j \leq \frac{j}{m.c(m)}\alpha$$

Avec c(m) mesurant le niveau d'indépendance, tel que:

$$c(m)=\sum_{i=1}^{m}\frac{1}{i}$$

1.5. Le séquençage du génome du blé

Le génome hexaploïde du blé tendre (17 Giga-base, soit 136 fois celui d'*Arabidopsis thaliana* ~125Mb) composé de trois génomes complets A, B et D a finalement été séquencé dans sa globalité mais l'assemblage reste brouillon (IWGSC, 2014). Ce travail est la réalisation de plusieurs groupes de chercheurs dans le monde qui ont travaillé dans le cadre d'un Consortium pour le Séquençage du génome du blé (IWGSC pour « International Wheat Genomic Sequencing Consortium »).

Chacun de ces trois génomes est composé d'environ 5500 millions de bases (http://www.wheatgenome.info). Les premières analyses ont permis d'identifier plus de 120 000 gènes. Toujours dans le cadre du consortium, le séquençage du chromosome 3B le plus long du génome du blé a été réalisé selon une méthode plus précise pour obtenir une carte physique plus fine.

L'IWGSC a mené une approche fondée sur les chromosomes et bras de chromosome pour la cartographie physique et le séquençage du génome du blé. Ce projet de séquençage reposent sur deux grands principes de séquençage des génomes (Barroy-Hubler, 2003), la première basé sur une méthode hiérarchique tandis que la deuxième demeure une méthode globale.

La publication de cette ébauche génétique est, une étape majeure vers l'obtention d'une séquence de référence du génome du blé tendre, l'objectif ultime du Consortium.

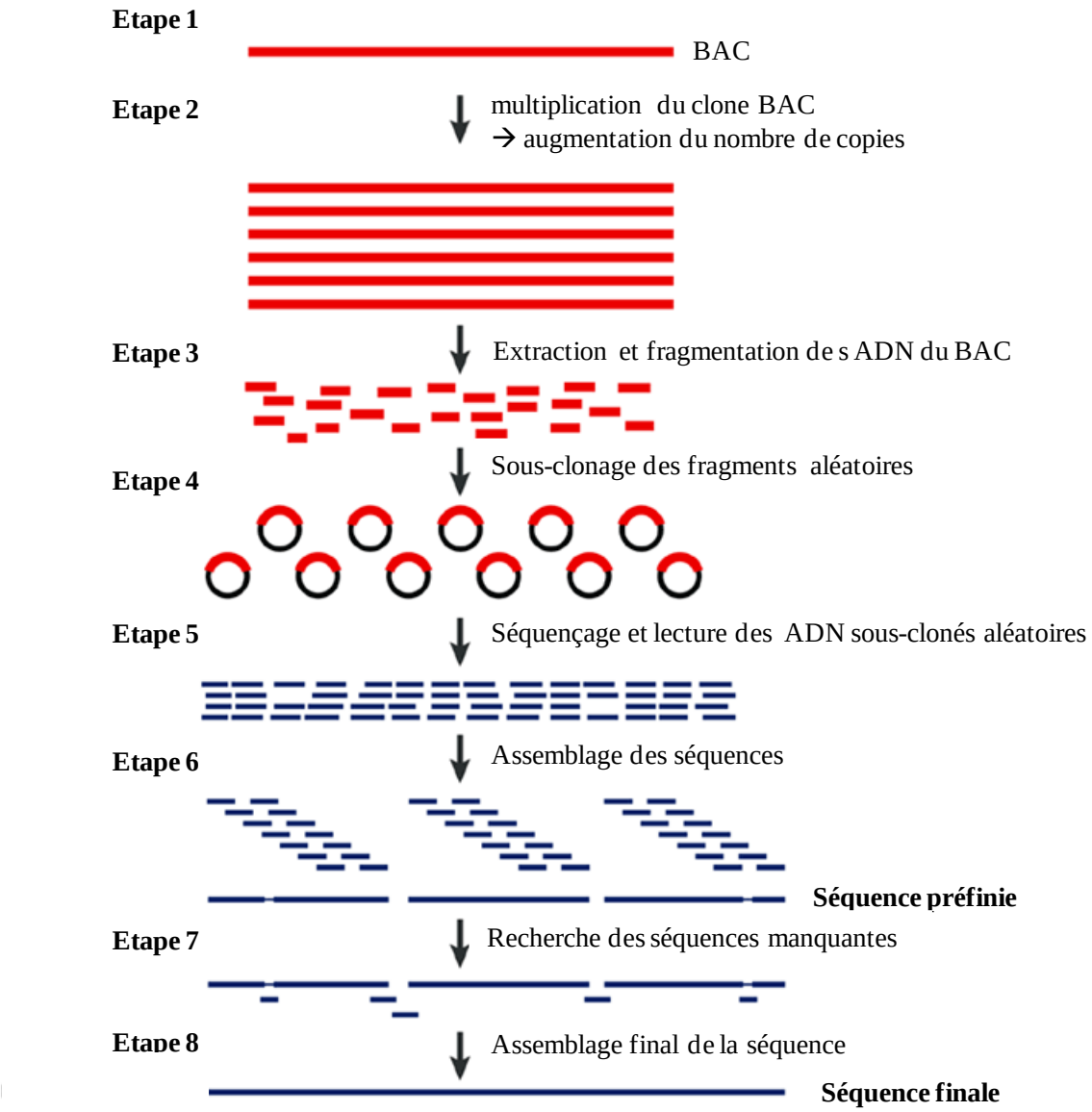


Figure 38: Etapes principales du séquençage clone par clone.

(1) Un clone BAC est sélectionné, et multiplié (2), une grande quantité d'ADN du BAC est purifié (comme lorsqu'on fait plusieurs photocopies d'une page spécifique d'un volume d'une encyclopédie). L'ADN purifié est fragmenté (les pages photocopiées passent dans un destructeur de papier) (3). Les fragments d'ADN aléatoires (2-5kb de taille) sont sous-clonés (4). Les lectures des séquences sont alors générées à partir d'une des deux extrémités des sous-clones choisis aléatoirement (des milliers de lectures ou « reads » sont générées pour chaque sous-clone) (5). Les lectures aléatoires sont assemblées sur la base du chevauchement des séquences (6), générant un ensemble de séquences préliminaires (séquence préfinie). Cette séquence présente des interruptions ou « gaps », et des régions où la séquence est de faible qualité (7). Pour finaliser la séquence, des données supplémentaires sont générées par d'autres méthodes permettant de compléter les interruptions et les régions de basse qualité, générant ainsi une séquence de très haute qualité (8). D'après (Green, 2001).

1.5.1. La méthode hiérarchique (séquençage clone par clone)

Le génome du blé hexaploïde a été cloné en banque BAC (bacterial artificial chromosomes) (Bresson et al., 2011). Les clones BAC sont des bactéries (*E.coli*) dans lesquelles sont stockées des molécules circulaires super-enroulées (plasmides) à l'état natif pouvant intégrer des fragments d'ADN de taille variant de 100 à 300kb issus de l'organisme à étudier (Shizuya et al., 1992). Le système repose sur un contrôle strict du nombre de copies de plasmides (ou vecteurs) dans la bactérie : soit une à deux copies par cellule. Ce nombre faible de copies par bactérie permet de limiter l'apparition de fragments d'ADN chimère (5%) par rapport à l'utilisation de banque YAC (Yeast artificial chromosomes, 20%). Le chimérisme est issu de plusieurs échanges d'ADN entre copies de plasmides dans une même cellule.

Avec les nouvelles méthodes de séquençage, la petite taille du vecteur BAC devient un critère de choix. Pour que le séquençage soit performant, le nombre de clones doit permettre une couverture de 5 à 10 fois la taille totale du génome étudié. Une fois référencé dans la banque, les clones sont fragmentés et séquencés individuellement. Puis les séquences sont assemblées par alignement en choisissant le recouvrement minimum ou « minimum tilling path », qui correspond au plus petit sous-ensemble de BAC chevauchants. Cette couverture est la plus complète possible du génome. (Fig. 38).

Les avantages de cette méthode sont :

- ✓ la grande facilité d'assemblage des fragments grâce aux chevauchements des BAC,
- ✓ la possibilité de comparer les fragments aux banques de données disponibles,
- ✓ la possibilité de répartir les régions chromosomiques pour le séquençage entre plusieurs laboratoires associés.

L'inconvénient majeur de la méthode est la difficulté de cloner des fragments contenant des séquences répétées très fréquentes notamment chez le blé (Barroy-Hubler, 2003).

Le premier séquençage d'un chromosome de blé, le chromosome 3B (le plus grand des chromosomes du blé, car à lui seul il représente 995 mégabases (soit 10 fois le génome d'*Arabidopsis thaliana*), a démarré par son clonage en banque BAC en 2004 (Safar et al. 2004) et vient d'être accompli (Choulet et al., 2014). Différentes approches ont été utilisées, le séquençage hiérarchique de clone BAC, mais également les technologies de nouvelle génération de séquençage (NGS) (Choulet et al., 2014). Au final, 8452 BACs ont été séquencés et assemblés pour obtenir une macromolécule d'ADN de 774 Mb composée à 85% d'éléments transposables (TEs), avec 5326 gènes codant pour des protéines et 1938 pseudogènes prédits.

La densité des gènes augmente le long de l'axe centromère-télomère, allant de 1.3 à 27.9 gènes par Mb (Choulet et al., 2014). Le séquençage du chromosome 3B représente une couverture total de 50x (50 fois sa taille), grâce à la combinaison de deux méthodes. La première est le séquençage hiérarchique des BAC en 454 Titanium et la deuxième méthode consiste en l'utilisation de l'analyseur Illumina GAIIx avec des lectures pairées de 600bp. L'analyse de la distribution : des sites de recombinaison, de la densité des gènes et des TEs révèle trois régions le long du chromosome 3B (Choulet et al., 2014):

- deux régions distales (~60Mb) qui présentent la plus forte densité en gènes et où la quasi-totalité des recombinaisons se sont déroulées ;

- une région proximale (centromérique) où la densité des gènes est la plus faible et les évènements de recombinaison quasi-absent.

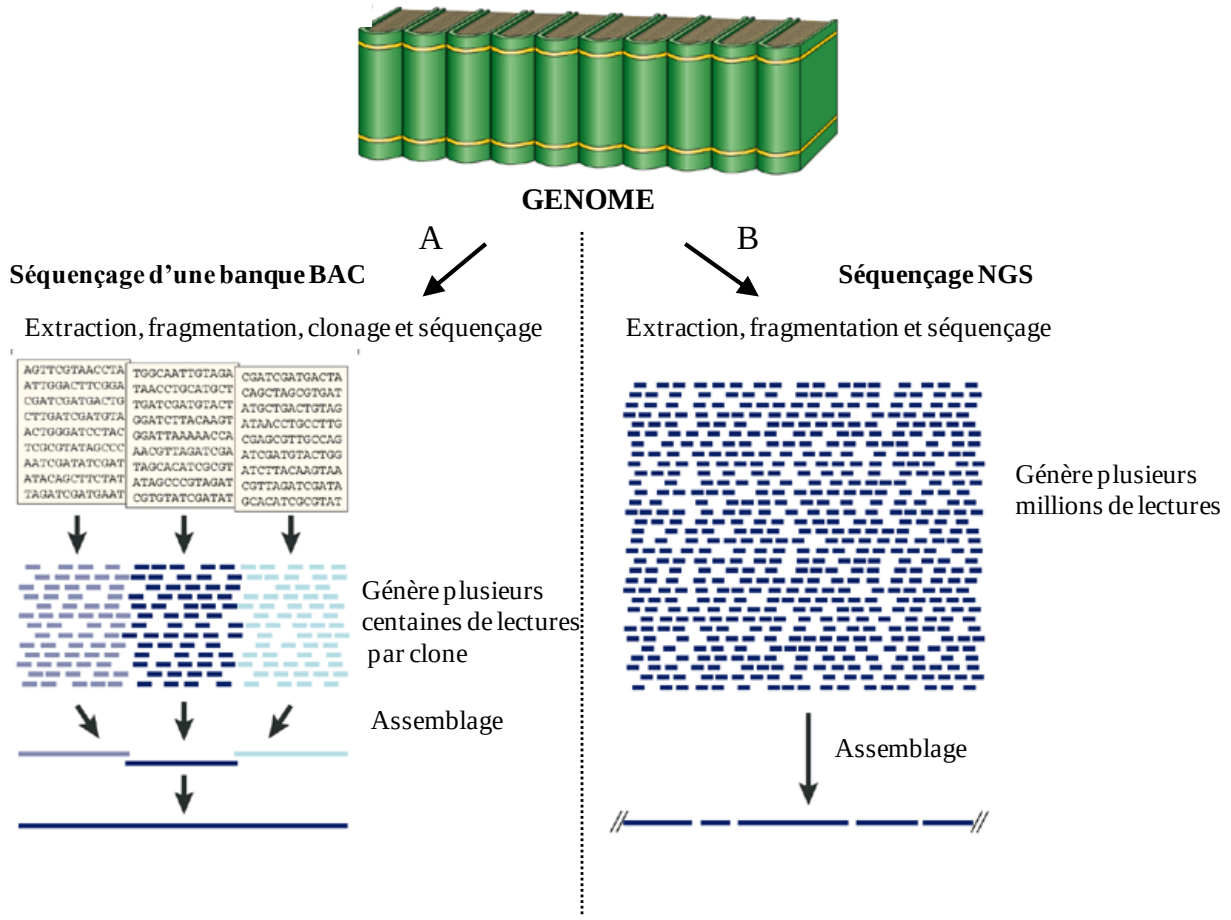


Figure 39: Les deux principales stratégies de séquençage en shotgun.

Vue d'ensemble du séquençage, si on représente le génome comme une encyclopédie, chaque volume représente un chromosome individuel. Il existe principalement deux stratégies pour obtenir une carte physique du génome.
A) La première stratégie consiste à extraire l'ADN génomique, le fragmenter et le cloner dans des bactéries (construction d'une banque BAC) ; puis dans une deuxième étape à extraire ces fragments, les séquencer et reconstituer la séquence initiale par le chevauchement des lectures. Chaque clone (BAC) contient un fragment d'ADN représentant une page d'un volume de l'encyclopedie. Le nombre de clones d'une banque dépend de la taille de génome.
B) La deuxième stratégie est le séquençage par « whole-genome shotgun ». Cette méthode passe par l'extraction, la fragmentation d'une grande quantité d'ADN génomique purifié (une librairie). Le séquençage est effectué en utilisant un nombre de librairies proportionnel à la taille du génome. Au final, 10 millions de lectures sont générées et analysées pour chaque librairie. L'assemblage des lectures d'une à plusieurs librairies, pour reconstituer la séquence du génome, est l'étape clef de la méthode. D'après (Green, 2001).

1.5.2. La méthode globale (ou whole-genome shotgun)

La méthode globale (ou whole-genome shotgun) séquence les fragments génomiques obtenus dans un ordre aléatoire, puis les réordonne par chevauchements. Cette méthode de séquençage a été initialement mise au point sur les génomes bactériens (Staden, 1979), le génome de la drosophile et enfin sur le génome de l'homme et celui de la souris. Elle a été récemment utilisée pour le séquençage du génome du colza hautement dupliqué (Chalhoub et al., 2014). Deux à trois banques composées de fragments d'ADN aléatoires de tailles différentes sont réalisées. À partir de ces banques, de nombreux fragments sont séquencés puis assemblés. La séquence totale est obtenue à force de recouvrement et d'assemblage.

Les principaux avantages de la méthode globale par rapport à la méthode de séquençage hiérarchique sont la rapidité de la mise en œuvre et le faible coût de la technique. Par contre un inconvénient majeur apparait lors du traitement informatique qui ne permet pas l'alignement de longs fragments de séquences répétées, ce qui est un problème chez le blé, car les séquences d'éléments répétés représentent 85% du génome.

Les principales différences entre ces deux stratégies sont le nombre d'étapes intermédiaires. En effet, l'ordonnancement hiérarchique passe par la fabrication d'une banque BAC, méthode délicate et coûteuse dans sa globalité, mais plus efficace pour l'alignement et le chevauchement des séquences répétées. A l'inverse, pour la méthode globale le génome entier réduit en fragments de petite taille est directement séquencé, mais l'inconvénient demeure la nécessité de puissante machines de calculs et la complexité d'assemblage des lectures de séquences répétées (Fig. 39).

En utilisant la technique de pyroséquençage Roche 454 sur des séquences issues de banques d'ADN génomique fragmentées ou « whole-genome shotgun », Brenchley et al. (Brenchley et al., 2012) ont identifié entre 94 000 et 96 000 gènes du blé allohexaploïde, et assignés les 2/3 aux génomes A, B ou D (Brenchley et al., 2012). Leur travaux ont aboutit à la création de familles de gènes orthologues et à la création d'une base de donnée de SNP[17] spécifiant pour chacun des gènes la position des SNP permettant de spécifier les homéologues.

[17] single nucleotide polymorphisms

Toujours dans le cadre du Consortium de Séquençage du Génome du Blé (IWGSC), un certain nombre de projets ont été réalisés, dont celui de Schreiber et al (Schreiber et al., 2012), qui a développé une approche basée sur les lectures Roche 454 et les lectures paired-end Illumina pour l'assemblage des transcrits des copies homéologues.

Tanaka et al (Tanaka et al., 2013) ont séquencé le chromosome 6B du génome du blé, à partir d'ADN génomique fragmenté ou « shotgun » des bras chromosomiques 6BS et 6BL. Ils ont assemblé, respectivement, 235 et 273 Mb pour couvrir ~55.6 et 54.9% des régions génomiques. L'assemblage des lectures montre respectivement, 77 et 86% de séquences répétées sur les bras chromosomiques 6BS et 6BL.

Parmi ces projets on notera l'utilisation de la stratégie du séquençage génomique aléatoire ou « whole-genome shotgun » sur un génome parental du blé.

En effet, Jia et al (Jia et al., 2013) ont séquencé et assemblé une ébauche ou « draft » du génome d'*Ae. tauschii*, ancêtre du génome D, avec une couverture de lecture de 90x (398Gb de lectures) à partir de 45 librairies. Pour chacune d'entre-elles, les tailles des inserts varient entre 200 bp et 20 kb.

Luo et al (Luo et al., 2013) ont séquencé le même génome, progéniteur du génome D du blé hexaploïde, pour établir une carte physique à partir d'une carte génétique à l'aide de 7 185 marqueurs ancrés sur les contigs totalisant une carte de 4.03 Gb. Par la suite, ils ont utilisé la méthode du séquençage aléatoire ou « whole genome shotgun », pour compléter et annoter la séquence du génome D. Le bilan de cette annotation met en évidence 17 093 gènes ou fragments de gènes (Luo et al., 2013).

Une approche similaire des projets réalisés sur le génome D a également été entrepris sur le progéniteur du génome A (*T. urartu*) par l'équipe de Ling et al.(Ling et al., 2013). Ils trouvent au final de leurs annotations 34 879 gènes ou fragments de gènes

Les cartes de diversité des génomes (genome-wide diversity maps) ont été élaborées récemment chez le blé allohexaploïde (Chao et al., 2009; Allen et al., 2011 ; Lai et al., 2012 ; Winfield et al., 2012; Allen et al., 2013 ; Cavanagh et al., 2013), chez les tétraploïdes (Saintenac, 2011; Trebbi et al., 2011; Ren et al., 2013) et les progéniteurs diploïdes (You et al., 2011; Wang et al., 2013).

L'assemblage préliminaire ou « draft » des 17 Gb génome du blé allohexaploïde (*Triticum aestivum*) a été réalisé très récemment (IWGSC, 2014). C'est un « premier jet » de séquençage des 21 chromosomes du génome du blé dans le sens où tous les gènes ont été identifiés sur le chromosome correspondant. Toutefois, l'orientation des gènes et les séquences intergéniques restent à préciser. Parmi les 124 201 séquences identifiées comme gènes ou loci de gène, 56 113 ont été positionnées le long des chromosomes de chaque sous-génome A B et D (soit respectivement : 17 297, 18 607 et 20 209 gènes) (IWGSC, 2014), et 8605 gènes représentent les trois homéoallèles des génomes A, B et D assignés aux différents bras de chromosomes. Ces triplets d'homéoallèles sont particulièrement intéressants pour l'étude de l'expression et des interactions entre homéoallèles.

Ces efforts de séquençage ont été répartis sur les différents bras de chromosome isolé. La méthode du séquençage aléatoire en masse ou « whole genome shotgun » a été utilisée pour identifier les gènes de blés diploïdes et tétraploïdes à l'origine du blé allohexaploïde. Les technologies NGS ont été utilisées avec une couverture de 30x (chacune des bases est couverte au moins 30 fois). Les séquences obtenues ont été assemblées en fragments, puis en contigs d'ADN unique, et ont été assignés à des positions le long des chromosomes par l'approche GenomeZipper (Mayer and al., 2009). Cette approche par collinéarité permet d'ordonner de manière virtuelle les gènes du blé en alignant ces contigs sur les cartes génétique du blé et d'autres espèces proches entièrement séquencées (*Brachypodium distachyon* (Initiative., 2010), le riz *Oryza sativa* (Project., 2005), le sorgho *Sorghum bicolor* (Paterson and al., 2009), et l'orge *Hordeum vulgare* (Mayer and al., 2012)).

1.6. Contexte et Objectifs des travaux

En utilisant pour modèle les espèces de blé (*Triticum* et *Aegilops*), l'objectif de ma thèse est de comprendre la dynamique complexe qui suit la formation de blé allopolyploïdes. Je me suis focalisée sur la caractérisation de la reprogrammation de l'expression des gènes et l'analyse de la contribution des homéologues.

Pour cet objectif, j'ai pu bénéficier tout d'abord des travaux menés au laboratoire (Chague et al., 2010) sur des données de puces Affymetrix, sur lesquelles je me suis appuyée pour aboutir aux travaux de (Chelaifa et al., 2013). Dans le but d'étudier l'expression des gènes dupliquées (homéologues) de façon séparée, j'ai développé par la suite une approche « Parent Spécifique » fondée sur les travaux d'Akhunova (Akhunova et al., 2010) et consistant à identifier les sondes d'oligonucléotides hybridant spécifiquement avec un des parents du blé hexaploïde et non avec l'autre, ce qui revient à élaborer une composante parent-spécifique de la puce Affymetrix.

Les technologies NGS/RNA-Seq offrent une meilleure résolution et permettent d'étudier plus finement l'expression des homéologues, de façon séparée. Le développement de ces technologies entraîne une évolution des outils d'étude de l'expression des gènes. Par conséquent, il était devenu important que je puisse me servir des données de séquençage mRNA-seq (Illumina), produites au sein du laboratoire. Ces données de séquençage ont abouti au développement d'outils d'analyse de l'expression des gènes homéologues pour un polyploïde dont la séquence référence du génome n'a pu être disponible qu'en fin de thèse. Les différents modèles d'études de l'expression des homéologues ont été élaboré dans une expérience pilote très enrichissante que je présente dans mon manuscrit.

Mes travaux de thèse se positionnent dans le cadre du projet ANR-Blanc 2011 Ploid-Ploid Wheat, dont l'enjeu majeur est d'éclaircir les réponses à la polyploïdie et à l'aneuploïdie chez les plantes à fleurs, en utilisant le blé pour modèle.

Deuxième Partie

Chapitre 2

Analyse des changements de l’expression des gènes dans les blés allohexaploïdes

2.1. Contexte et questions posées

La polyploïdie est un mécanisme majeur d'évolution des espèces, déclenchant la réorganisation des génomes au niveau génétique, fonctionnel et épigénétique (Madlung and Wendel, 2013 ; Soltis and al., 2013).

Les premières caractérisations des effets de la polyploïdie au niveau fonctionnel ont été réalisées, principalement, par comparaison entre les allopolyploïdes synthétiques (produits en condition de laboratoire) et/ou les allopolyploïdes naturels et leurs progéniteurs (Jackson and Chen, 2010), ou encore la moyenne parentale dite MPV (Mid Parent Value) (Pumphrey et al., 2009 ; Chague et al., 2010 ; Chelaifa et al., 2013 ; Zhang et al., 2014). Comme décrit dans l'introduction générale, la comparaison à la MPV permet de montrer ou non l'hypothèse d'additivité. Ainsi, l'expression d'un gène est considérée comme additive dans le polyploïde lorsqu'elle ne diffère pas de l'expression moyenne ou de la somme de l'expression des gènes parentaux. L'expression d'un gène est dite non-additive, lorsqu'elle en diffère.

Les caractérisations des changements de l'expression des gènes dans les blés hexaploïdes ont été faites dans mon laboratoire, en utilisant les puces Affymetrix GeneChip Wheat GenomeArray qui comportent 61 178 sondes (ou probesets). Chacun de ces « probesets » est constitué de 11 oligo-sondes de 25 nucléotides, dessinés sur 61 178 transcrits différents du blé (Fig. 23). Les données disponibles dès le début de mes travaux de thèse, m'ont permis d'apprendre et d'appliquer la méthode d'analyse des données micro-array Affymetrix. Ainsi, nous avons comparé les changements d'expression des gènes dans l'allohexaploïde synthétique du blé $B_hA_hD_t$ et ses progéniteurs.

Les premières caractérisations des allohexaploïdes synthétiques du blé, ayant une stabilité moyenne au niveau de la régularité de la méiose, ont permis de caractériser 7% de gènes à expression non additive (Chague et al., 2010). Néanmoins des allohexaploïdes synthétiques beaucoup plus stables ont été également caractérisés (Mestiri et al., 2010). Ces polyploïdes ont comme donneur du génome AB l'allotétraploïde « Tetra-Courtot », extrait à partir de l'allohexaploïde naturel cv Courtot (Mestiri et al., 2010). Il est devenu donc assez important de caractériser comparativement la reprogrammation de l'expression des gènes. Ce travail a fait l'objet d'une publication dans la revue *New Phytologist* (Chelaifa et al., 2013).

Avant de présenter cet article, je décris ici la procédure d'analyse et de comparaison de l'expression des gènes utilisée dans ces travaux.

2.1.1. Le prétraitement des données

Les données ont été normalisées par la méthode gcRMA (Irizarry et al., 2003), décrite dans le chapitre 1.

2.1.2. L'analyse différentielle

2.1.2.1. Evaluation des variations intra-génération

Afin de considérer deux plantes sœurs d'une même génération comme réplicats biologiques, les variations intra-génération ont été évaluées par **ANOVA** (ANalysis Of Variance).

L'analyse de la variance à un facteur permet de comparer les moyennes d'une variable donnée dans k populations à partir d'échantillons aléatoires et indépendants, prélevés dans chacune d'elles.

La variance totale, la variance intragroupe et la variance intergroupe sont analysées pour comparer les moyennes et tester l'hypothèse H_0: $\mu_1 = \mu_2 = ... = \mu_k$

L'ANOVA suppose que :

- la variable étudiée suit une distribution normale ;
- les variances des populations sont toutes égales ;
- et les échantillons sont tirés aléatoirement et indépendamment dans les populations.

L'ANOVA consiste à valider l'homogénéité des variances intra et inter classes pour comparer les moyennes des échantillons.

Après avoir validé les plantes sœurs comme réplicats biologiques, les niveaux d'expression de gènes entre les différents génotypes ont été comparés.

2.1.2.2. Les gènes différentiellement exprimés

Pour comparer les différences d'expressions entre les gènes, la modèle **VarMixt** (Delmar et al., 2005) a été utilisé. C'est un modèle de mélange sur les variances permettant de repérer les

gènes différentiellement exprimés entre deux conditions. Ce modèle repose sur l'hypothèse que les groupes de gènes peuvent être identifiés sur la base de réponse similaire aux différentes sources de variabilité. La variance de chaque groupe peut-être estimée précisément à partir d'un grand nombre d'observations.

En utilisant les variances de groupe, la statistique de test pour un gène g devient :

$$t_g^{\mathrm{VM}} = \frac{\bar{y}_{g1\cdot} - \bar{y}_{g2\cdot}}{S_g^{\mathrm{VM}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où :

$\bar{y}_g$ est la moyenne d'expression du gène g, sur l'ensemble des réplicats, dans chaque condition ;

n est la taille de l'échantillon dans chaque condition ;

S_g^{VM} est l'écart type du gène g estimé par VarMixt:

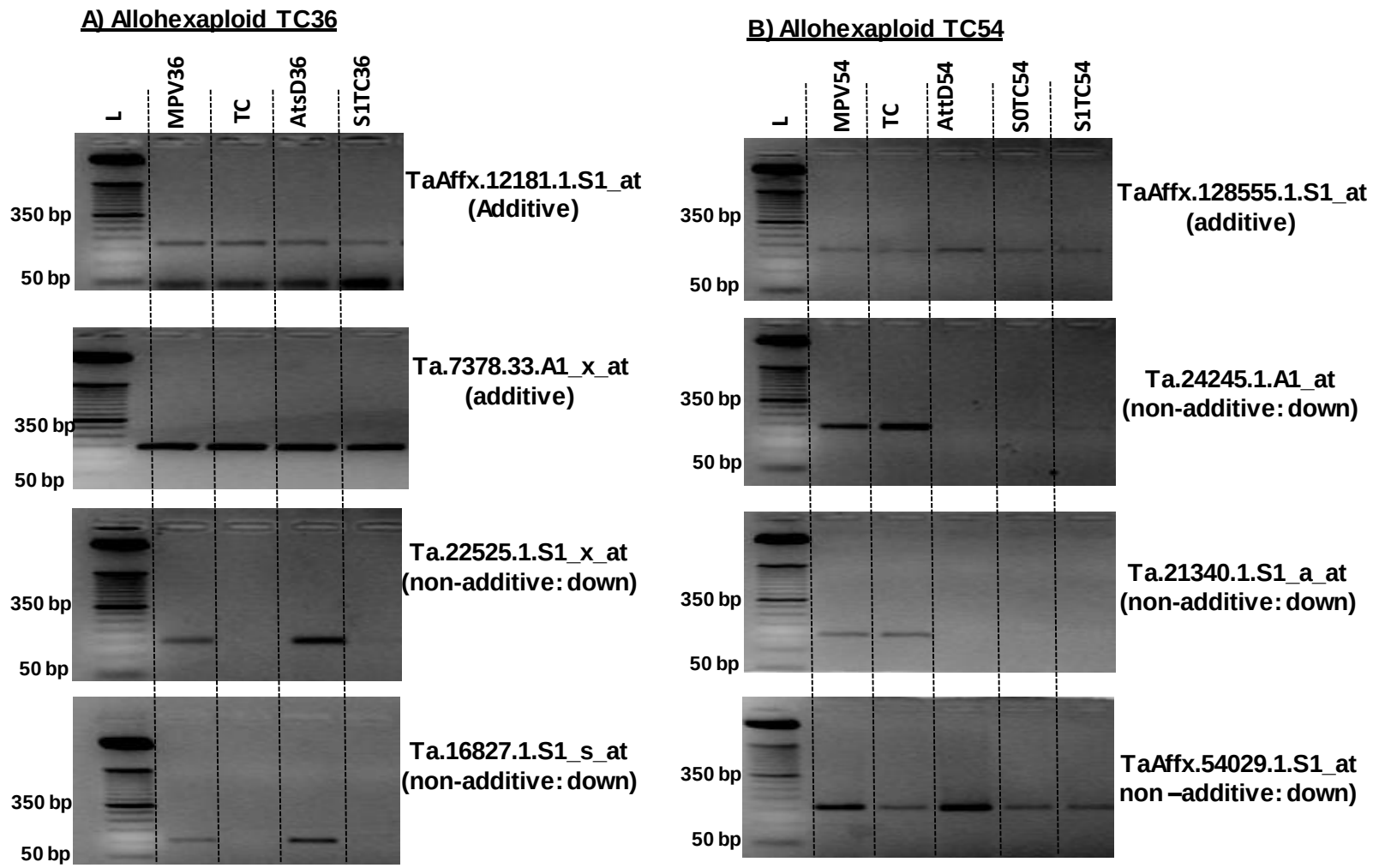
$$S_g^{\mathrm{VM}} = \sqrt{\sum_1^k \pi_{gi} S_{G_i}^2}$$

avec G_1, ..G_k les k groupes de variances, π_{gi} la probabilité d'appartenance du gène g au groupe i, et S^2_{Gi} la variance du groupe G_i.

La variance d'un gène est donc une somme pondérée des variances de groupes.

2.2. Prévalence de l'additivité de l’expression des gènes dans des blés allohexaploïdes génétiquement stables.

Cette partie est présentée sous forme d’article, qui a été publié dans la revue *New Phytologist*.



Supporting Information Fig. S1. Confirmation by semi-quantitative RT-PCR of expression pattern of eight genes revealingconsistency with microarrays data in the wheat synthetic allohexaploids TC36 (A) and TC54 (B). The synthetic allohexaploids TC36 and TC54 were obtained through hybridization between the extracted tetraploid Tetra-Courtot and two accession of *Ae.tauschii (*AtsD36 & AttD54) followed by spontaneous chromosome doubling. Primers used for RT-PCR are the same as described in Chagué *et al.* (2010).
MPV: mid-parent values. L : 50 bp Ladder (*NewEngland Biolabs*), 50 and 350 bp size are indicated.

Supporting Information Table S2 Comparison of frequencies of biological process gene ontology (GO) terms in sets of non-additively expressed genes revealed in TC36 and/or TC54 wheat synthetic allohexaploids.* significant enrichment at *P* value < 5%

GO category	Genes on µarray	Non-additive genes in at least one allohexaploid					
		Up regulated		Down regulated		Total	
	Occurrence	Occurrence	Enrichment	Occurrence	Enrichment	Occurrence	Enrichment
-Metabolic process	7,827	16	0.14	36	0.33	52	0.48
Photosynthesis	190	0	0	2	0.76	2	0.76
-Cellular process	7,164	9	0.09	11	0.11	20	0.22
Transcription	765	0	0	1	0.09	1	0.10
-Biological regulation	1,345	2	0.10	9	0.48	11	0.59
Regulation of transcription	721	1	0.10	2	0.20	3	0.30
-Reproduction	**46**	**3**	**4.74***	**0**	**0**	**3**	**4.74***
-Cellular component organization or biogenesis	638	4	0.45	6	0.68	10	1.14
-Localization	1,496	3	0.14	6	0.29	9	0.43
-Developmental process	154	0	0	0	0	0	0.00
-Response to stimulus	**659**	**5**	**0.55**	**20**	**2.2***	**25**	**2.76***
-Cell death	130	0	0	0	0	0	0.00
-Multi-organism process	32	0	0	0	0	0	0.00
-Multicellular organismal process	34	0	0	0	0	0	0.00
		0					
-No annotation	15,628	90	0.41	66	0.30	156	0.73
-No match	21,903	102	0.31	185	0.61	222	0.74

En raison de la dimension du tableau S1, en données supplémentaires, je vous propose de vous référer directement au tableau en ligne sur :
http://onlinelibrary.wiley.com/doi/10.1111/nph.12108/suppinfo

Chapitre 3

Changements de l'expression des gènes en diminuant et ré-augmentant le niveau de ploïdie chez le blé:

Analyse de l'expression globale et de l'expression parent-spécifique

3.1. Contexte et questions posées

Les changements rapides et dynamiques dans l'expression des gènes des plantes polyploïdes, reflètent la plasticité du génome (Jackson and Chen, 2010). Cette plasticité dans la régulation des gènes dupliqués, dont une grande partie provient de facteurs épigénétiques, offre une nouvelle voie pour la diversification fonctionnelle et l'évolution des traits adaptatifs chez les polyploïdes (Jackson and Chen, 2010). La plasticité du génome est un facteur clé dans le succès des espèces polyploïdes, notamment chez le blé (Dubcovsky and Dvorak, 2007).

Plusieurs études ont porté sur l'expression des homéologues dans les allopolyploïdes pour un nombre limité et ciblé de gènes (Koh et al., 2010; Buggs et al., 2011a; Dong and Adams, 2011; Combes et al., 2012), ou au niveau du génome entier (Bancroft et al., 2011 ; Higgins et al., 2012 ; Combes et al., 2013; Chalhoub et al., 2014). L'hybridation et la duplication du génome mènent à la divergence de l'expression des homéologues chez le coton (Yoo et al., 2013).

Afin de disséquer la contribution des génomes des progéniteurs dans le blé allohexaploïde, j'ai développé et affiné une méthode originale qui utilise l'outil Affymetrix GeneChip Wheat GenomeArray (utilisé dans le chapitre précédant) pour distinguer l'expression des sous-génomes AB et D. Ces travaux ont été réalisés sur la base des travaux d'Akhunova et al. (Akhunova et al., 2010). Ainsi, on utilise la technologie Affymetrix GeneChip Wheat GenomeArray à l'échelle des 11 « oligos-sondes » composant chaque 'probeset' (Fig. 23) et on identifie ceux qui s'hybrident à un parent (donneurs du génome AB ou D) et pas à l'autre (sondes « parents-spécifiques »).

Afin d'identifier ces sondes Parent-Spécifique (ou « Parent-Specific Feature », noté PSF), j'ai utilisé une instance du modèle linéaire présenté dans les travaux d'Akhunova (Akhunova et al., 2010). Dans cette étude, toutes les sondes PM sont ajustées au modèle linéaire ci-dessous:

$$Y_{ijn} = \mu + \text{sonde}_i + \text{génotype}_j + \varepsilon_{ijn}$$

avec :

Y_{ijn}, le niveau d'expression du transcrit du réplicat n (n allant de 1 à 4) du génotype j pour la sonde i ;

μ, l'intensité moyenne de toutes les sondes (indépendante de i, j, n) ;

sonde $_i$, l'effet de la sonde i (avec i allant de 1 à 11) ;

génotype $_j$, l'effet du génotype j (avec j étant le génotype AB ou D) ;

ε_{ijn}, les résidus contenant l'effet d'intéraction sonde-génotype.

L'analyse différentielle est réalisée sur les résidus ε_{ijn} contenant l'effet d'interaction entre la sonde et le génotype, afin d'identifier les sondes hybridant différentiellement un des génotypes. Pour d'identifier ces sondes parent-spécifiques (ou sous-génomes spécifique AB ou D), la méthode Limma (Linear Models for Microarray Data) a été utilisée (Smyth, 2004). Cette méthode modélise la variance de la différence d'expression (Martin-Magniette, 2012), une distribution *a priori* de la variance est utilisée sur l'ensemble des observations (ou gènes) pour estimer une loi *a posteriori* de la variance du gène. La méthode Limma utilise un nombre de degré de liberté supérieur à celui d'un t-test classique, ce qui permet d'améliorer la puissance du test (Martin-Magniette, 2012).

La statistique de test limma est :

$$t_g^{\text{limma}} = \frac{\bar{y}_{g1\cdot} - \bar{y}_{g2\cdot}}{S_g^{\text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où :

$\bar{y}_g$ est la moyenne d'expression du gène g (sur l'ensemble des réplicats) dans chacune des conditions (condition 1 et condition 2) ;

n est la taille de l'échantillon dans chaque condition ;

S_g^{limma} est l'écart type du gène g :

$$S_g^{\text{limma}} = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}$$

avec:

S_0^2 la variance *a priori* ;

d_0 le nombre de degrés de liberté *a priori* ;

S_g^2 la variance du gène g sur les observations des conditions n_1 et n_2 ;

d_g est le nombre de degrés de liberté empirique (n_1+n_2 - 2) du gène g.

La distribution *a priori* sur les variances permet d'emprunter de l'information à l'ensemble des gènes pour mieux en déduire celle de chacun des gènes individuellement.

3.2. Changements de l’expression des gènes lors de la diminution et de la réaugmentation de l’alloploidie dans le blé.

Cette partie est présentée sous forme d’article, non-encore soumis.

Unraveling gene expression changes when decreasing and re-increasing alloploidy in wheat

Chalabi Smahane[1], Chelaifa Houda[1], Dominique Arnaud[1], LeFloch Edith[2], Mestiri Imen[1], Vinh Ha DinhThi[1], Isabelle LeClainche[1], Claudine Devauchelle[2], Denise Deffains[3], Virginie Huteau[3], Olivier Coriton[3], Harry Belcram[1], Carène Rizzon[2], Julien Chiquet[2], Joseph Jahier[3], and Boulos Chalhoub[1*].

[1]Unité de Recherche en Génomique Végétale URGV (INRA-CNRS – UEVE), Organization and Evolution of Plant Genomes, 91057,Evry Cedex, France;

[2]Laboratoire Statistique et Génome, Université d'Evry Val d'Essonne, UMR CNRS 8071 - USC INRA, Evry, France

[3]Unité Mixte de Recherches INRA, Agrocampus Rennes - Université Rennes 1, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), 35653, Le Rheu, France;

[*]Corresponding author email: chalhoub@evry.inra.fr

Abstract

- To improve our understanding of reprogramming of gene expression in response to allopolyploidy, we used the original wheat models where ploidy level can be either: (i) decreased by extracting the allotetraploid component B_hA_h from the natural hexaploid wheat *Triticum aestivum* ($B_hA_hD_h$) and also; (ii) re-increased through synthetic allopolyploidization between the extracted B_hA_h allotetraploid and the D_t genome donor diploid species *Ae. tauschii*.
- Various comparisons of gene expression changes have been performed at the global gene expression level, using the Affymetrix GeneChip Wheat Genome Array; and its partitioning between B_hA_h and $D_{h/t}$ subgenomes for 514 genes, for which we developed parent specific features ("PSF").
- Global gene expression comparisons reveal that the majority of 34 820 transcripts are equally expressed, with only 497 ones (1.4%) that show significant expression differences between natural hexaploid and its extracted tetraploid wheat. Interestingly, 83.5% of these readopt, when adding again the D_t genome in a newly-synthesized allohexaploid, similar expression to natural allohexaploid. Sixty-one out of the 70 remaining genes are not (or are down) expressed in *Ae. tauschii*, explaining thus the reason for not re-adopting, in synthetic allohexaploids, equal expression to natural ones. We were able to analysis subgenome contribution for 514 genes which reveal that B_hA_h and $D_{h/t}$ subgenomes are expressed in allohexaploid wheats at generally 2/3 and 1/3 of their expression levels in their respective progenitor species and also partition global gene expression by 2/3 for the A_hB_h and 1/3 for the $D_{t/h}$ subgenomes, respectively.
- Our study brings the unprecedented scope of 'minimal' global gene expression changes when decreasing and/or re-increasing ploidy levels in wheat, and reveals massive subgenome expression compensation and concerted partitioning, in accordance with the balance in gene dosage.

Introduction

Polyploidy, or the occurrence of more than two sets of chromosomes in one single nucleus, is an important and recurring process in the evolution of angiosperms (Wendel, 2000; Osborn et al., 2003 ; Adams, 2007; Chen, 2007 ; Doyle et al., 2008 ; Tang et al., 2008 ; Soltis and Soltis, 2009 ; Van de Peer et al., 2009a ; Chen, 2013; Madlung and Wendel, 2013). Two fundamental types of polyploids are known: autopolyploids, which derive from duplication of same genome of a same species and allopolyploids, where two or more divergent homoeologous genomes became united by interspecific or intergeneric hybridization followed by chromosome doubling.

During the last 15 years an increasing number of studies have shown that polyploids undergo changes at the genetic level (Song et al., 1995 ; Ozkan et al., 2001 ; Rieseberg, 2001; Shaked et al., 2001; Gaeta et al., 2007 ; Mestiri et al., 2010; Chalhoub et al., 2014) as well as at the gene expression and/or epigenetic levels (Kashkush et al., 2003; Adams and Wendel, 2005b; Wang et 2006; Flagel et al., 2008 ; Hovav et al., 2008 ; Ha et al., 2009b ; Pang et al., 2009 ; Rapp et al., 2009 ; Chague et al., 2010 ; Grover et al., 2012 ; Qi et al., 2012 ; Chelaifa et al., 2013 ; Yoo et al., 2013 ; Guan et al., 2014). The extent and the 'timing' of these changes depend on the analyzed allopolyploid but natural allopolyploids rarely correspond to a simple addition of progenitor genomes (Comai, 2005 ; Chen, 2007 ; Leitch and Leitch, 2008 ; Soltis and Soltis, 2009 ; Arnaud et al., 2013).

Ubiquity of polyploidy in plants may be largely due to the propensity of angiosperms to undergo chromosomal duplication and subsequent DNA rearrangements, known as the diploidization process, but the derived diploid never resembles to any of its progenitor species (Paterson et al., 2004; Adams and Wendel, 2005 ; 2005b ; Liu et al., 2009 ; Soltis and Soltis, 2009). Recent studies have indicated that transcriptomic changes in allopolyploids may be an adaptive mechanism that facilitates the establishment and the evolution of a stable species (Paterson et al., 2011 ; Rambani et al., 2014 ; Xu et al., 2014).

Wheat species provide a good example of relatively recent and stable allopolyploids. The widely cultivated allohexaploid wheat *T. aestivum ssp. aestivum* (2*n*=6*x*=42, BBAADD), also known as common or bread wheat, originated as the result of two separate amphiploidization events. The tetraploid *T. turgidum* ssp.*dicoccum* (2*n*=4*x*=28, BBAA) arose less than 0.5 million years ago as a result of hybridization between *T. urartu* Tumanian ex Gandylian (2*n*=2*x*=14, AA) and an unidentified diploid *Aegilops* species of the section *Sitopsis*, thought to be *Ae. speltoides* (2*n*=2*x*=14, SS) or a close relative thereof, as the donor

of the B genome (Riley et al., 1958 ; Dvorak and Zhang, 1990 ; Dvorak et al., 1993 ; Takumi et al., 1993 ; Talbert et al., 1995; Blake et al., 1999; Huang et al., 2002b ; Chalupska et al., 2008). A spontaneous hybridization between the early-domesticated tetraploid *T. turgidum* ssp. *dicoccum* and the diploid goatgrass *Ae. tauschii* (2*n*=2*x*=14, DD), about 10 000 years ago, gave rise to *T. aestivum* (Kihara, 1944 ; McFadden and Sears, 1946 ; Nesbitt and Samuel, 1996; Huang et al., 2002b).

At the genomic level, we recently showed that except for variation in homologous pairing, leading to chromosome instability and aneuploidy, no DNA sequence elimination or other rearrangements are observed when analyzing euploid plants of newly-synthesized allohexaploids (Mestiri et al., 2010 ; Zhang et al., 2013). The apparent additivity observed at the structural level (Mestiri et al., 2010), could be largely attributed to the *Ph1* (*Pairing homoeologous 1*) locus, which enhances the exclusive pairing of homologous chromosomes at metaphase I of meiosis and constitutes the main stabilizing factor of these wheat polyploids (Riley and Chapman, 1958; Griffiths et al., 2006), preventing thus homoeologous recombination (Mestiri et al., 2010). The apparent additivity of BA and D genomes in hexaploid wheat has been early exploited in order to extract BA tetraploids from hexaploid wheat, thus eliminating the D genome, through recurrent backcrossing and cytogenetic characterization (Mestiri et al., 2010 ; Zhang et al., 2013). Our recent characterizations have shown that synthetic wheat allohexaploids having extracted tetraploids as AB genome progenitor are more stable than those having a natural wheat tetraploid, in term of complete homologous pairing and low aneuploidy frequency (Mestiri et al., 2010).

An increasing number of studies, including wheat, have characterized effects of increasing alloploidy level in natural or synthetic allopolyploids and shown variable proportions of genes which expression deviates from the average of their progenitors (between 0.3% to 19%, depending on the species and studies) (Adams and Wendel, 2004 ; Hegarty et al., 2005; Wang et al., 2006b; Pumphrey et al., 2009; Rapp et al., 2009; Akhunova et al., 2010; Chague et al., 2010; Chelaifa et al., 2013).

In the present study, we exploit the possibility of extracting AB tetraploids from natural allohexaploid wheat, as well as resynthesizing allohexaploid wheat, as done previously , in order to characterize effects of both decreasing and then re-increasing allopolyploidy levels on gene expression regulation. As detailed in Mestiri et al. (2010), it is possible to extract from the allohexaploid wheat cv. Courtot (having the A_h, B_h and D_h genomes, subscript denotes homoeologs in the allohexaploids) the tetraploid A_hB_h component

‘Tetra-Courtot’, through recurrent backcrossing and cytological characterization, thus removing the D_h genome (Fig.1). The D genome (from *Ae. tauschii* accessions*:* AttD54, and AtsD36, having D_t genome subscript denotes *tauschii*) is added to the extracted B_hA_h tetraploid ‘Tetra-Courtot’, through synthetic allohexaploidization, thus allowing re-increasing ploidy level (Fig.1, detailed in (Mestiri et al., 2010)). In this original model, it was possible to perform various comparisons that allow us to characterize interaction between B_hA_h and D_h subgenomes, gene expression stability in allohexaploid wheat and the evolution of novel gene expression regulation.

Materials and Methods

Synthetic wheat allopolyploids material and growth conditions

The same euploid plants characterized by Mestiri *et al.* (2010) as not displaying structural changes, were analyzed for gene expression regulation as compared to their progenitors as well as to mid-parent values (MPV). All plants of progenitors and synthetic allohexaploids as well as the natural wheat allohexaploid (*T. aestivum*) cv. Courtot were grown in growth chambers at 22°C and 16 hours day length.

Gene expression was measured in two sister plants, considered as biological replicates, for each of the analyzed genotypes: *T. turgidum spp. durum* cv Joyau (B_jA_j, 2n=4x=28, j subscript denotes Joyau), the goatgrass *Ae. tauschii* accessions AttD54 and AtsD36 (D_t, 2n=2x=14) as described in (Mestiri et al., 2010), the natural allohexaploid cv. Courtot and cv. Chinese Spring ($B_hA_hD_h$, 2n=6x=42), their respective extracted tetraploid components Tetra-Courtot and Tetra-Chinese Spring (B_hA_h, 2n=4x=28) (Fig.1, detailed in (Yang et al., 1999; Mestiri et al., 2010)).

For analysis of transcriptome changes, we chose in this study the synthetic allohexaploids TC36 and TC54 having extracted tetraploid Tetra-Courtot as AB genome donor and *Ae. tauschii spp. stangulata* accession 36 (AtsD36) and *Ae. tauschii spp. tauschii* accession 54 (AttD54) as D genome donors, respectively (Mestiri et al., 2010). We compared the first-selfed (S0) generation of TC54, the second generation (S1) of TC36, and an *in vitro* mixture of equal amounts of RNAs from progenitors to measure the MPVs as described in (Chague et al., 2010; Wang et al, 2006).

This resulted in 22 samples that were hybridized and analysed.

Affymetrix GeneChip Wheat Genome Array hybridization

To analyze gene expression changes in wheat leaves, we used the available Affymetrix GeneChip® Wheat Genome Array. The array contains 61 178 probesets representing 55 049 transcripts for all 42 chromosomes in the wheat genome (http://www.affymetrix.com/). Total RNA extraction, reverse transcription, synthesis of the double stranded cDNA, in vitro transcription of cRNA, labeling and hybridization on the Affymetrix GeneChip® Wheat microarray were done as described by (Chague et al., 2010). Arrays were scanned with the GeneChip® Scanner 3000 7G piloted by the GeneChip® Operating Software (GCOS). All these steps were performed on Affymetrix platform at URGV laboratory, Evry, France.

- **Standard Probeset level analysis**

The raw '.CEL' files were imported into the R software and the 22 arrays were altogether normalized with gcRMA as in (Chague et al., 2010). The plant material considered for the analysis at the probeset level is detailed in Table 1. For comparisons between the different genotypes, transcripts were considered as expressed (detected) in a given genotype when values of hybridization intensity (I), expressed as log-ratio after normalization, were higher than '3' (I>3), the Affymetrix detection cutoff, in both biological replicates. If one or both replicates are inferior or equal to '3' (I≤3), the transcript was considered as not detected. This resulted in a dataset of 34 820 probesets, representing ~57% of the total number of probesets printed on the array (61 178) that has been compared between the different conditions and genotypes. The Supporting Information, Table S1, indicates the overall number of expressed transcripts for all genotypes and all replicates analyzed in the present study.

All statistical analysis were done as in Chagué et al. (2010) according to the varmixt model that relies on identifying groups of genes having homogeneous variance (Delmar et al., 2005).

We compared the hexaploid wheat cultivars Courtot and Chinese Spring to their B_hA_h extracted tetraploid "Tetra-Courtot" (TC) and Tetra Chinese Spring (TCS), where the D_h has been removed, in order to characterize the effect of decreasing the allopolyploidy level. We also performed a variety of comparison and cross comparisons between these polyploid wheats and the D_t genome donor *Ae. tauschii* (accessions AttD54 and/ or AtsD36), the natural wheat allotetraploid *T. turgidum spp durum* cv Joyau

In order to characterize the re-increasing allopolyploidy level, we analyzed gene expression changes in newly-synthesized wheat allohexaploids obtained between the B_hA_h tetraploid "Tetra-Courtot" and the D_t genome progenitor *Ae. tauschii* by conducting various comparisons with natural wheat allohexaploids and progenitors. While the B_hA_h genome of Tetra-Courtot is nearly the same as the one of the allohexaploid wheat cv Courtot, the exact D_h genome progenitor of the natural allohexaploid wheat does not exist anymore and cannot be extracted from the allohexaploid wheat. Thus, we analyzed two synthetic allohexaploids TC36 and TC54 that were obtained by hybridizing, followed by spontaneous chromosome doubling, the same B_hA_h extracted tetraploid Tetra-Courtot with two different accessions of *Ae. tauschii* (AtsD36 and AttD54, respectively), to better represent D genome variability. We focused on characterizing how those genes which expression was shown as changing when

eliminating the D_h genome, i.e. differentially expressed between Courtot and "Tetra-Courtot", are behaving when adding again the D_t genome in synthetic wheat allohexaploids.

- **Parent Specific Features (PSF) level analysis**

We also used Affymetrix expression array to discover parent-specific oligo-probes as described in Akhunova et al. (2010) with more stringent modifications as described below. The plant material used in this section is detailed in Table 1. The microarrays Affymetrix Genechip Genome Array contains 61 178 probesets, each with 11 probes, thus ending with 672 958 oligo features (probes). Parent Specific features (hereafter PSF) are identified as the oligonucleotides showing statistically stronger hybridization intensity for one of the parental genome using the Affymetrix data at the oligo-probe level. As AB representative, we combined expression data from the natural allotetraploid Joyau and the extracted tetraploid "Tetra-Courtot" for the robustness of identified PSFs. For the D genome, we used *Ae. tauschii* accessions AttD54 and AtsD36. We performed on the associated CEL the usual background correction and quantile normalization (RMA normalization). This is done using the affy R-package available on Bioconductor.

For detection of PSF, we rely on the principle of PSF discovery (Rostoks et al., 2005; Akhunova et al., 2010) and Limma procedure (moderated t-test) (Smyth, 2004) which is known to show more statistical power than SAM (Jeanmougin et al., 2010). We apply the multi-test correction of Benjamini and Hochberg (Benjamini and Hochberg, 1995) on the Limma p-values. Thus, we can decipher probes which are PSFAB or PSFD based on this statistic and its associated adjusted p-value, with a threshold to $\alpha = 5\%$. We also remove PSF which belong to a probeset with strictly more than 3 PSF (Akhunova et al., 2010) (Fig.2).

We refined the set of PSF obtained via the above procedure by removing the PSF specific to a given parent AB or D but that show significant interfering hybridization with the opposite parent. For such an assessment, we compared the hybridization of a given PSF in an equivalent-molar mixture of the two parental RNA (MPV) to that of its hybridization in the parent where it is specific (parent AB or D). A PSF is considered as interfering if its expression level in the mixture is significantly higher than ½ of the expression level measured in its specific parent. The arrays used to perform this test are background corrected, quantile normalized and log2 transformed. We performed a classical Welch-test with Benjamini and Hochberg correction on the p-values, with a threshold to $\alpha = 5\%$ (Fig.3). By

default, we performed Limma test for all comparisons, but we could not use it here for one-sided tests and we use a Welch-test instead.

Among non-interfering PSF, we removed PSF whose hybridization level in the genome where it is assumed to be not to be specific is significantly more than 3, and kept PSF whose hybridization level in the genome where it is assumed to be specific is significantly more than 3, as recommended by Affymetrix. For all these filters, we perform a Welch-test with the Benjamini and Hochberg's correction and set the threshold to $\alpha = 5\%$. These filters led to 3 types of probesets, those with only PSFAB, named Probeset AB genomes or PSTAB, those with only PSFD, named PSTD and those PSFAB and PSFD, named PSTABD. Table 2 details statistics obtained through each filter.

In our work, we focus on the 514 elaborated PSTABD to assess B_hA_h and D_t subgenome contribution to global gene expression. For probesets composed of two PSF specific of a same parent, the corresponding PSF is the mean of the expression level of this two PSF. Under the assumption of probesets PSTABD, we defined, for a gene, global gene expression as the sum of B_hA_h homoeoallele and D homoeoallele expression levels.

To understand effects of decreasing allopolyploidy level, we compared PSF expression level of our set of 514 genes between natural allohexaploids $B_hA_hD_h$ (combining Courtot and Chinese Spring) and corresponding B_hA_h extracted tetraploids (Tetra-Courtot and Tetra-Chinese Spring, as detailed in Table 1). We also compared natural allohexaploids to the D genome donor diploid species, represented here by AttD54 and AtsD36. Then to understand the effect of re-increasing allopolyploidy level, we compared allohexaploid synthetics, TC54 and TC36, to their genome progenitors B_hA_h extracted tetraploid and D_t diploid species. We also assessed the number of additive genes to MPV in allohexaploid synthetics. For these comparisons, we considered global gene expression as defined previously and PSF expression level, we performed Limma tests correction with the Benjamini and Hochberg's correction and set the threshold to $\alpha = 5\%$.

Results

The particular experimental scheme and the wheat allopolyploid material were designed in this study in order to characterize effects of both decreasing and then re-increasing allopolyploidy levels, by respectively eliminating and then adding again the D parental genome (Fig.1). As it was described since the sixties (Kerber, 1964), it was possible to re-extract from the hexaploid wheat cv. Courtot (having the A_h, B_h and D_h genomes) the tetraploid B_hA_h component "Tetra-Courtot", through recurrent backcrossing and cytological characterization (Fig.1, (Mestiri et al., 2010 ; Chelaifa et al., 2013). It is important to precise that, after six recurrent backcrosses, the B_hA_h genome of Tetra-Courtot was shown to be ~100% identical to the B_hA_h genome of Courtot when 838 mapped markers were compared ((Mestiri et al., 2010) and Unpublished). Thus, this experiment could be assimilated to decreasing the allopolyploidy level where the D_h genome was eliminated (Fig. 1, detailed in (Mestiri et al., 2010). The D_t genome, from two accessions of *Ae. tauschii* (AttD54 and AtsD36), was added again to the extracted B_hA_h tetraploid "Tetra-Courtot", through synthetic allohexaploidization, thus allowing re-increasing the ploidy level (Fig.1, detailed in (Mestiri et al., 2010).

Gene expression changes when decreasing the allopolyploidy level

- **Global gene (Probeset) expression profile**
    - **comparing allohexaploid wheat and extracted B_hA_h tetraploids**

The majority of the considered set of 34 820 transcripts are equally expressed with only 497 transcripts (1.4%) that show significant expression differences ($p<0.05$) between Courtot and Tetra-Courtot. The majority of these differentially expressed transcripts (457~1.3%) are up-expressed in the hexaploid wheat cv Courtot, whereas (40~0.1%) transcripts are down-expressed (Fig.4c). It is most-likely that majority of the 457 transcripts that are down-expressed in 'Tetra-Courtot' have been contributed in the allohexaploid wheat cv. Courtot by the D_h genome. Similarly, we can also hypothesize that the 40 transcripts that are up-expressed in Tetra-Courtot would have been repressed in the allohexaploid wheat cv. Courtot by the D_h genome. Alternatively, one can also suggest that some of the differentially expressed transcripts B_hA_h would have evolved or adopted a different gene expression-regulation in the natural allohexaploid wheat (than its progenitors) and this could be or not maintained in its extracted tetraploid 'Tetra-Courtot'. While the exact AB and D genome

progenitors of natural wheat allohexaploid are no more available, a primary way to better understand these different hypotheses is to analyse gene expression in other extracted wheat allotetraploids as well as natural wheat allotetraploid.

We characterized and compared gene expression changes in another extracted tetraploid "Tetra Chinese Spring" (hereafter called TCS) (Kerber, 1964), as compared to its progenitor hexaploid wheat cultivar "Chinese Spring" (CS), as well as the natural wheat allotetraploid *T. durum* cv. Joyau. Cross comparison between the two extracted tetraploids (TC and TCS) and their corresponding progenitor hexaploid wheat cultivars (C and CS, respectively) (Fig.5c-d), showed a higher proportion (3.47%) of differentially expressed genes between Chinese Spring and its extracted tetraploid (877~2.52% up-expressed and 333~0.95% down-expressed). However, majority of genes differentially expressed between the natural allohexaploid cv Courtot and its extracted tetraploid Tetra Cortout (342/457~74.8% up-expressed and 11/40~27.5% down-expressed) were common to those differentially expressed between CS and TCS (Fig.5c-d).

Interestingly, the two extracted wheat allotetraploids showed a lower proportion of differentially expressed genes (Fig.5g) between each others, with 116 and 170 genes respectively up-expressed in TC as compared to TCS. Both extracted wheat allotetraploids, TCS and TC exhibit a higher proportion of up-expressed genes as compared to the natural allotetraploid *T. durum* cv Joyau (Fig.5a-f), but there were more differentially expressed genes between TCS and Joyau (551 up- and 221 down-expressed) than between Joyau and TC (366 up- and 211 down-expressed) (Fig.5). However, majority of the later (226/366 and 77/211) were common to those revealed when comparing TCS to Joyau (Fig.5a-h). On the other hand, majority of the remaining genes, differentially expressed between the natural allotetraploid *T. durum cv Joyau* and one or the two extracted tetraploids, were equally expressed between these and their corresponding allohexaploid progenitors, Chinese Spring and Courtot (Fig.5a-f)

Further comparisons also shows that majority of genes that were down expressed in Tetra-Courtot as compared to Courtot were similarly down-expressed in *T. durum* cv Joyau as compared to Courtot (429/457) (Fig.5b-c). Majority of these were equally expressed between Tetra-Courtot and *T. durum* cv Joyau (423, Fig.5b-c). Similarly majority of genes that are down-expressed in TCS as compared to CS (605/877) were also down expressed in *T. durum* cv Joyau (Fig.5d-e), Majority of these were equally expressed between TCS and *T. durum* cv Joyau (597, Fig.5d-e).

- **comparing allohexaploid wheat and the D genome donor diploid species**

Another way to better elucidate the different hypotheses is to analyse gene expression, and particularly those genes that are differentially-expressed between the hexaploid wheat and their extracted allotetraploids in the natural diploid wheat *Ae tauschii* (two accessions Att54 and Ats36 were used). We used here those genes reveled between Courtot and Tetra-Courtot as they show the least gene expression differences.

The set of transcripts that were differentially up- or down- expressed in Courtot as compared to 'Tetra-Courtot' are also compared for their expression between 'Tetra-Courtot' and *Ae. tauschii* accession Ats36. Majority of the 457 transcripts that are up-expressed in Courtot, as compared to 'Tetra-Courtot' are also up-expressed in Ats36 compared to 'Tetra-Courtot' (365, representing 79.8%), suggestion that the up-expression in Courtot was most-likely contributed by the D genome in Courtot (Fig.4cd). Similarly, 34 transcripts that are down-expressed in Courtot as compared to Tetra-Courtot are also down-expressed in Ats36 (Fig.4cd). Similar results are obtained when comparing transcript expression between 'Tetra-courtot and the other accession of *Ae. tauschii* AttD54 (Table S2).

- **Parent Specific Feature (PSF) comparisons**

In order to dissect expression and resolve portioning of B_hA_h and D subgenomes, we focused on analysis of expression of the 514 probesets for which we revealed both PSFAB and PSFD (PSTABD) (Table 1, See Material and Method). In PSF expression comparisons, we considered natural allohexaploid wheats cvs Courtot and Chinese Spring together, for extracted allotetraploids TC and TCS, for the diploid D genome donor *Ae. tauschii* AttD54 and AtsD36 to better represent the intra species variability (Table 1).

- **Comparison of natural allohexaploid $B_hA_hD_h$ *vs* extracted tetraploids B_hA_h**

Considering global gene expression level measured here as the sum of composing PSF in each genotype, we confirm that majority of genes are expressed at equal levels 460/514 (89.5%) between natural allohexaploids $B_hA_hD_h$ (considering CS and C together) and the extracted tetraploid (considering TC and TCS together) (Table 3.a). For half of these equally expressed genes 224/460~48.7%, the two PSF A_hB_h and D, are not significantly differentially expressed in the natural allohexaploid $B_hA_hD_h$, and for almost the other half of

genes 218/460~47.4% B_hA_h homoeoalleles are up-expressed compared to D_h PSF. Only 3.9% of genes similarly expressed in natural allohexaploids and their extracted tetraploids exhibit an up-expression of D_h homoeologs in natural allohexaploid.

The up-expression of B_hA_h PSF, in natural allohexaploid, is most likely measuring expression of A_h and B_h homoeologs that can have different partitioning in comparison to each other and to the D_h homoeolog. Each of the three homoeologs can be less or equally expressed to the other, but A_h and B_h could not be distinguished, as measured together (PSFAB). This could be better precised when A_h and B_h could be resolved and their expression could be separately evaluated such as through RNA-Seq technologies. This balanced homoeolog bias towards the sum of A_h and B_h homoeologs expression in allohexaploid wheat that we are revealing here is also in concordance with earlier findings showing that B_hA_h genome in hexaploid wheat contributes two parts and the diploid parent contributes only one part to total gene expression in the allohexaploid (Akhunova et al., 2010). This is also validated by the fact that, in natural allohexaploid, for more than half of genes (285/514 ~ 55.4%) where 1/2 of B_hA_h homoeologs expression and the D_h homoeolog expression are not, significantly, different. For the remaining genes, (137/514)~26.6% of genes exhibit an up-expresssion of ½ of B_hA_h homoeologs expression level compared to D_h homoeolog and only 92/514~17.9% genes are up-expressed for D_h homoeolog (Table 3.b).

Among the (53+1/514)~10.5% genes differentially expressed, at the global level, between the natural allohexaploids wheats and the extracted tetraploids, we also confirm that the majority (53/54~98.1%) are up-expressed in the natural allohexaploids (Table 3a). Most of these (44/53~83%) having a higher expression level of D_h subgenome; compared to B_hA_h subgenome. Only a few genes exhibit an equal expression between B_hA_h and D_h subgenomes in natural allohexaploids (7/53~13.2%) and two genes (2/53~3.8% =) exibit up-expression of B_hA_h subgenomes.

- **Comparison of natural allohexaploids *vs* B_hA_h tetraploids wheat and *Ae. tauschii***

Interestingly, we observed a diminution of subgenome expression level in natural allohexaploids compared to extracted tetraploids B_hA_h and D diploid species (Fig.6a). The distribution of the ratio of B_hA_h homoeologs expression level in natural hexaploid ($B_hA_hD_h$) over their expression in the extracted B_hA_h exhibits a unique and well-defined peak of density equal to 2/3 (Fig.6c). In natural allohexaploids $B_hA_hD_h$, for all the 514 genes, B_hA_h

subgenome expression level does not, significantly, differ from 2/3 of their expression level in extracted tetraploids B_hA_h (Table 4).

Similarly, the distribution of the ratio, of D_h subgenome expression level in natural allohexaploids over their expression level in the D_t diploid species, exhibits also a peak around 1/3 (Fig.6c). Indeed, in natural allohexaploids, D_h homoeoalleles expression level does not significantly, differ for the majority of the genes (499/514~97%), from 1/3 of their expression level in the D diploid species (Table 5). Only few genes (12+3/514~3%) differ from 1/3 of the expression level in the D diploid species, most (12/15~80%) of these are significantly lower than 1/3 of D diploid genome expression level, the remaining 3 exhibit an up-expression in natural allohexaploids (Table 5).

- **Comparison of B_hA_h extracted tetraploids and the diploid D genome donor species *Ae. tauschii***

Comparing extracted B_hA_h tetraploids and the D genome donor diploid species, we observed that 245/514~47.6% of genes exhibit equivalent expression levels (Table 6). Among these, 148/245~60.4% exhibit similar expression level for B_hA_h and D_h subgenomes in the natural allohexaploids (Table 6.a) whereas 93/245~37.9% showed dominance expression of B_hA_h and only (4/245)~1.6% for the reverse situation (Table 6.a). Interestingly, 197/245~80.4% of these genes adopt the model of 2 to one contribution, where 1/2 of B_hA_h subgenome expression level equals D homoeologs level (Table 6.b) whereas for 35/245~ 14.3%, B_hA_h sugenome contributed more than two folds the D_h one.

For (58/131~44.2%) of genes more expressed in diploid species *Ae.tauschii* compared to extracted tetraploids, the expression level of D subgenome is dominating B_hA_h subgenome expression in allohexaploid wheat whereas the remaining genes (73/131)~55.7% are equally expressed; (53/131)~40% of these follow the model of 2 B_hA_h to 1 D_h contribution.

A higher gene expression level has been observed in extracted tetraploids B_hA_h compared to diploid species for 138/514 ~27% genes, most of these genes (104/138~75.4%) (Table 7) exhibit an expression dominance of B_hA_h subgenome in natural allohexaploids $B_hA_hD_h$, these genes are not significantly differentially expressed between natural allohexaploids and extracted tetraploids and are maintained up-expressed in tetraploids compared to D_t diploid species.

All together, in natural allohexaploids, we observe no significant difference, for more than half of genes (285/514~55.5%), when we compare ½ of B_hA_h homoeologs to total D_h

subgenomes expression level (Table 6.b). The remaining genes exhibit for (92/514) ~17.9% an up-expression of D_h subgenome (compared to ½ of B_hA_h subgenome expression level) and for (137/514)~26.6% an up-expression of ½ of B_hA_h subgenome expression level (Table 6.b).

Effect of re-increasing allopolyploidy level: Addition of D genome to the extracted B_hA_h tetraploid

- **Evaluation at the global gene expression level**

We performed here a number of cross-comparisons in order to characterize effects of re-increasing alloploidy level on the transcripts whose expression in the B_hA_h extracted tetraploid Tetra-Courtot was affected by the D_h genome elimination. For that, we sorted out transcripts differentially expressed between the natural allohexaploid cv Courtot and its B_hA_h tetraploid component Tetra-Courtot (Fig.4bc). We then analyzed their expression in newly synthesized allohexaploids, where the D_t genome of *Ae. tauschii* was added to the B_hA_h extracted tetraploid (Fig.1). As similar transcriptome changes were observed in the two synthetic allohexaploids TC54 and TC36, we present hereafter results observed for the allohexaploid synthetic TC36.

The comparison shows that 84.4% (386) of the 457 transcripts that became differentially up-expressed (in the original natural wheat allohexaploid cv Courtot compared to its extracted tetraploid TC) when the D_h genome was turned–off, i.e. when Tetra-Courtot was extracted from Courtot, re-adopt equal expression levels to the original natural wheat allohexaploid cv Courtot, after adding to Tetra-Courtot the D genome of *Ae. tauschii* acc. AtsD36 (in the newly synthetized allohexaploid TC36, Fig.4bc).

Among the remaining (71 transcripts), 70 are down-expressed in both the synthetic wheat allohexaploid TC36 and Tetra-Courtot as compared to the natural wheat allohexaploid Courtot (Fig.4bc). More detailed cross-comparisons show that 61 of these 70 transcripts are also down-regulated in the diploid progenitor Ats36 compared to natural allohexaploid Courtot (Fig 4ab, on the red dotted line). Thus, this category of transcripts may reflect differences between the two D genomes and reveal more precisely the specific higher contribution of the D_h genome progenitor of Courtot to gene expression in the natural allohexaploid.

- **Analysis of genes that have Parent Specific Features**

  - **Comparison synthetics vs MPV**

Considering global gene expression level, our results confirm the prevelance of additivity of parental gene expression in synthetic allohexaploids (TC36 and TC54) as compared to MPV (MPV-TC36 and MPV-TC54) (457/514~88.9%) (Table S3) (Chelaifa et al., 2013). Among these additive genes, more than half of them (243/457~53.2%), follow the model of 2 to one contribution of the B_hA_h expression and D homoeologs.

  - **Comparison of synthetics allohexpaloids with their progenitors**

As in natural allohexaploids, we observed a diminution of B_hA_h and Dt subgenome expression in synthetic allohexaploids compared to their progenitors (Fig.6b). The distribution of the ratio of B_hA_h subgenome expression level in synthetic allohexaploids to their expression level in the progenitor B_hA_h extracted tetraploid Tetra-Courtot reveals also a well-defined peak equals to 2/3 (Fig.6d). In synthetic allohexaploids, for almost the 514 genes (507/514, ~98.6%), B_hA_h subgenome contribution does not, significantly, differ from 2/3 of their expression level in the progenitor B_hA_h, with only 3 genes above and 4 below the 2/3 of the B_hA_h expression in the tetraploid progenitor (Table 8).

The distribution of the ratio of D_t subgenome expression level in synthetic allohexaploid to their expression level in the D genome progenitor exhibits also a peak around 1/3 (Fig.6d). It confirms that, in synthetic allohexaploids, the D_t subgenome contributes, for the majority of the genes (463/514~90%), to 1/3 of their expression level in the D genome progenitor (Table 9). We observe, however, 10% of genes that diverged from this 1/3 model, with 36 genes (7%) above and 15 genes (3%) are below to 1/3 of the expression of D homoeologs in the diploid D genome progenitor.

  - **Contribution of the two genomes in synthetic hexaploid wheat**

For the (245/514~47.6%) genes not differentially expressed, between the two progenitors, most are partitioned in the newly synthetized allohexaploid following the model 2:1 B_hA_h to D_t subgenome contributions (1/2 of homoeologs B_hA_h expression level equals D_t homoeologs level in the newly synthetized allohexaploid) (187/245~76.3%) (Table 10.b). For the remaining, (36/245)~14.7% expression in the newly synthesized allohexaploids is contributed

by more than 2/3 for B_hA_h subgenome and for 9% (22/245) by more than 1/3 by the D homoeolog (Table 10.b).

For the 131/514~25.5% genes more expressed in D diploids than B_hA_h genome progenitor, most of them (89/131~67.9%) maintain in the newly synthesized allohexaploids a higher expression of D_t homoeologs level than ½ of B_hA_h subgenome expression level (Table 10.b), and the remaining genes (42/131~32.1%) rebalance the subgenome expression level to the 1 to 2 partitioning model.

Among the 131 genes up-expressed in D diploid species, 34 exhibit an expression dominance of D_t subgenome in newly synthesized allohexaploids (Table 11). Crossing natural and newly-synthesized allohexaploids, we observe 20 genes commonly revealing an expression level dominance of D_t subgenome.

Majority of genes up-expressed in allotetraploids genome compared to diploids (105/138~76.1%), exhibit a dominant contribution of B_hA_h tetraploid subgenomes in newly synthesized allohexaploids (Table 11). Crossing the observations on natural and newly-synthesized allohexaploids reveals 88 genes sharing commonly an expression level dominance of B_hA_h subgenome.

Discussion

The allohexaploid wheat in which three divergent genomes became 'co-resident' in a single nucleus, through two allopolyploidization events, represents a fantastic model of a highly stable allopolyploid model where homoeologous recombination is prevented by the action of the *Ph1* gene considered as the main stabilizing factor (Riley and Chapman, 1958 ; Griffiths et al., 2006). At the genomic level, we have recently confirmed such stability and showed that except for variation in homologous pairing, leading to chromosome instability and aneuploidy, no DNA sequence elimination or other rearrangements are observed when reproducing the second allopolyploidy event in newly-synthesized allohexaploid wheat (Mestiri et al., 2010). This apparent structural stability, suggesting simple co-residence of AB and D genomes in allohexaploid wheat is confirmed by, and certainly the reason of success in extracting viable AB tetraploids from hexaploid wheat, where the D genome is eliminated Kerber (Kerber, 1964 ; Mestiri et al., 2010).

However, changes are observed at the gene expression level (Pumphrey et al., 2009 ; Akhunova et al., 2010, Chelaifa et al. 2013; Chague et al., 2010). These studies, as well as those on other plants, have always characterized effects of increasing alloploidy level (Adams and Wendel, 2004 ; Hegarty et al., 2005 ; Rapp and Wendel, 2005 ; Wang et al., 2006b ; Pumphrey et al., 2009 ; Akhunova et al., 2010; Chague et al., 2010). In the present study we exploited the possibility of extracting AB tetraploids from allohexaploid wheat, and thus elimination of the D genome, in order to characterize the effect of decreasing alloploidy level. It was then possible to add again the D genome and re-increase allopolyploidy levels, through synthetic allopolyploids. By analyzing natural allohexaploids ($B_hA_hD_h$), their extracted tetraploids (B_hA_h), and diploid D genome progenitors, our analysis allows detection of all genes which global expression is altered when eliminating the D_h genome. More interestingly, we checked for these genes, whose expression is altered by turning-off the D_h genome, their expression regulation when turning-on again the D genome, by re-increasing allopolyploidy level through synthetic allohexaploids. For this later analysis and in order to overcome the fact that the exact D_h genome progenitor of the natural allohexaploid wheat does not exist anymore, we used two distinct D genome progenitors (Ats36 and Att54) to better represent its variability when reincreasing ploidy level (synthetic allohexaploids).

Yet, we found that the majority of the transcripts have the same expression profiles when decreasing allopolyploidy level by eliminating the D genome. Only 497 transcripts (1.4%) are differentially expressed between the natural allohexaploid ($B_hA_hD_h$) and its

extracted tetraploid ($B_hB_hA_hA_h$), majority of which (411/497~83.5%) readopt equal expression when adding again the D genome in the newly synthesized allohexaploid TC36. A large majority of these genes (91.9%) (457/497) are down-expressed in the extracted tetraploid (B_hA_h) compared to the natural allohexaploid cv Courtot (Fig.4c). Among these, we found (70/457) genes that remain down-expressed in the newly synthesized allohexaploid TC36 compared to the natural (Fig.4.bc), majority of which (61/70) are interestingly also down-expressed in D genome progenitor (AtsD36) compared to the natural allohexaploid cv Courtot (Fig.4.abc). These differences could be, thus, inferred to a divergence in expression between the two different D_h and D_t genomes.

In our study, in addition to "Tetra-Courtot", we characterized and compared another extracted tetraploid "Tetra Chinese Spring" as well as the natural wheat allotetraploid *T. durum* cv. Joyau. Interestingly, the two extracted wheat allotetraploids showed a low proportion of differentially expressed genes between each others, whereas they were almost two times more differentially expressed genes between Joyau and TCS than between Joyau and TC (Fig.5a-f). The higher proportions of up-expressed genes in both extracted allotetraploids as compared to the closest known relative *T. durum* (Fig.5) confirm those of a recent study (Zhang et al., 2014) with a third different extracted wheat allotetraploid.

Cross comparison between the two extracted tetraploids and their corresponding progenitor hexaploid wheat cultivars (Fig.5), which has not been supplied by Zhang et al. (2014), showed a higher proportion of differentially expressed genes between Chinese Spring and its extracted tetraploid than between Courtot and its extracted tetraploid, majority of the later 342/457 and 11/40 were common to CS and TCS (Fig.5). Further comparisons also show that majority of genes that were down expressed in "Tetra-Courtot" as compared to Courtot were also the case for *T. durum* cv Joyau (429/457) (Fig.5b-c) and similarly for Chinese Spring as compared to TCS (Fig.5d-e). These findings suggest that other effects than evolution under allohexaploidization and selection under domestication (Zhang et al., 2014) such as the exact progenitor(s) of the AB subgenome of allohexaploid wheat and the extraction method and purity of TCS would also explain these observed differences.

In wheat, analysis of partitioning of global expression between homoeologs of a small set of genes had shown that homoeolog balance play a crucial role in the global gene expression. Using a synthetic allohexaploid, Stamati et al. (2009) showed that Thionins I expression pattern is contributed by the different homoeologous transcripts (Stamati et al.,

2009). Similarly, Nomura et al. (2005) studied the transcript levels and the catalytic properties of the three *TaBx* homoeologs in hexaploid wheat and showed homoeolog compensation between the three wheat genomes (Nomura et al., 2005). Zhang et al. (2011) analyzed the expression Q gene in polyploid wheat, and showed that the evolution of the *Q/q* loci in polyploid wheat resulted in the hyperfunctionalization of homoeoallele 5A*Q* that is more expressed than the two others, pseudogenization of homoeoallele 5B*q*, and subfunctionalization of homoeoallele5D*q* (Zhang et al., 2011). Recently, Hu et al. (2013) showed that expression divergence in TaEXPA1 homeologs in allohexaploid wheat is mediated by histone methylation and acetylation (Hu et al., 2013).

In our study, we used the Affymetrix GeneChip® Wheat Genome Array to derive parent specific oligoprobes and analyze partitioning of the global gene expression between the BA and D subgenomes. Parent Specific Features permit to estimate contributions of B_hA_h and D subgenomes in natural and newly-synthesized allohexaploids in comparison to their expression in tetraploid and diploid wheat species. A multiple model approach allowed us to study how allohexaploid wheat responds to the merging of genomes. Most of these genes, (Tables 5 and 9 respectively), follow in allohexaploid wheat the model of 1/3 of expression of D subgenome as compared to its expression in the diploid species (Fig.6c and 6d), and most of detected B_hA_h subgenome, follow the model of 2/3 of their expression level in the extracted B_hA_h tetraploid (Fig.6c and 6d, Tables 4 and 8). Moreover, the global gene expression in allohexaploid wheat was found to be 2/3 contributed by the A_hB_h subgenome and 1/3 for the D subgenome for majority of analyzed genes.

The scope of ‘minimal’ global gene expression changes observed here when eliminating the D_h genome, is unprecedented, and suggests that there is a massive subgenome expression compensation in this wheat polyploid model. Our results strongly suggest a dynamic process of gene expression regulation, involving “counterbalancing mechanisms” (Grover et al., 2008), a balance in dosage (Birchler et al., 2005), in regulatory and in stoichiometric relationships (Birchler and Veitia, 2007). Our study brings out genes expression evolution patterns involved in the fate of duplicated genes, in allohexaploid wheat. It reveals a high stability of the two genomes progenitor B_hA_h genome and D genome, at their coresidence, and confirms the establishment of gene expression regulation in early generations and its high stability upon evolution of wheat allohexaploid (Chague et al., 2010; Arnaud et al., 2013).

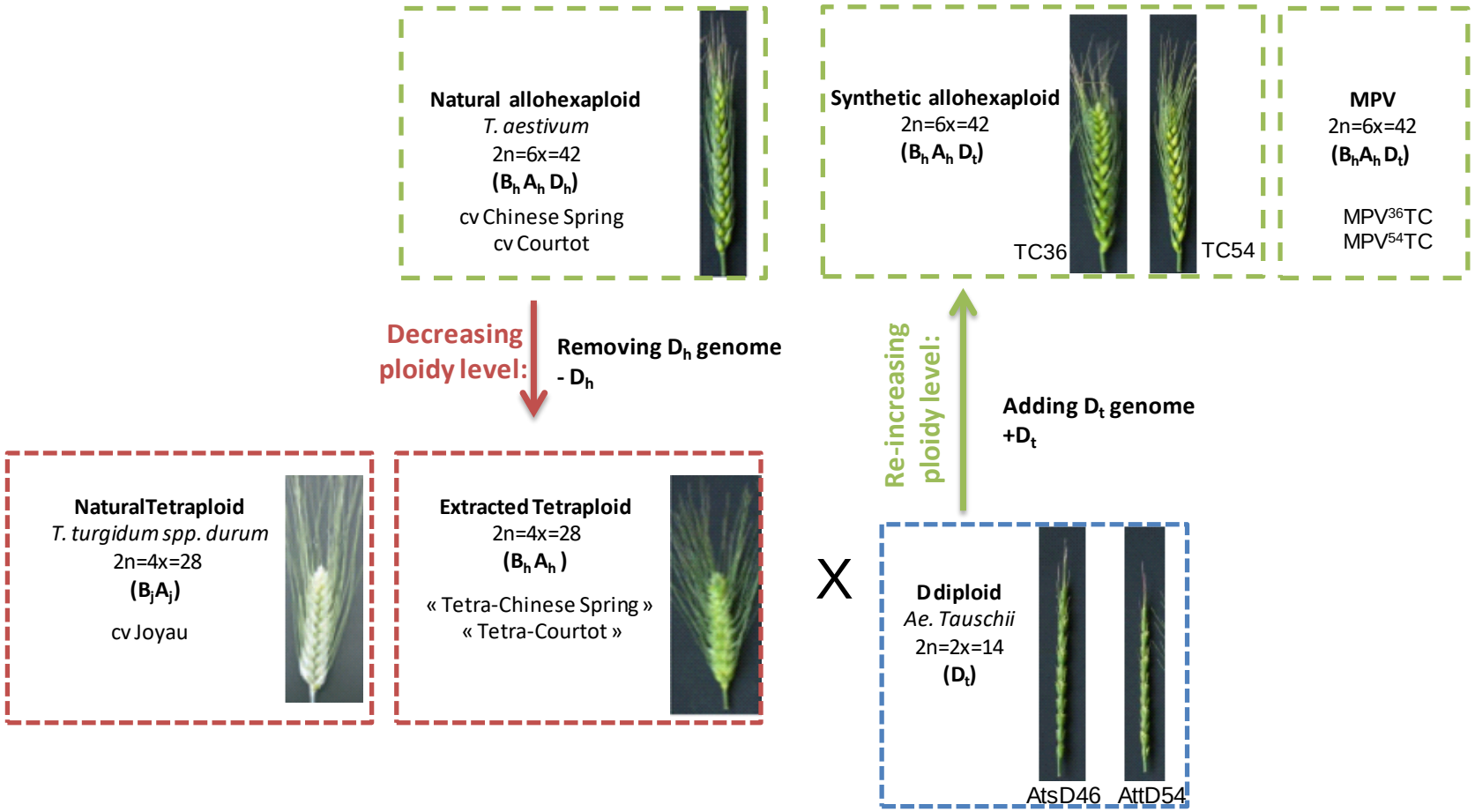


Figure 1: Schematic representation of origin of the extracted tetraploid, synthetic wheat allohexaploid and Mid-Parent Value (MPV) (details see Mestiri et al, 2010). The tetraploid component «Tetra-Courtot » and « Tetra-Chinese Spring »(2n=4x= 28, B_hA_h) are extracted from allohexaploid wheat cv Courtot and cv Chinese Spring, respectively, according to Kerber (1964)(Mestiri et al, 2010). Newly synthetized wheat allohexaploid were obtained through hybridization between « Tetra-Courtot » and *Aegilops tauschii* (2n=2x=14, D_t), followed by spontaneous chromosome doubling. Here, photos represente spike of the natural allohexaploid cv Courtot, the natural tetraploid *T.turgidum*, the extracted allotetraploid « Tetra-Courtot », D diploid specie *Ae.tauschii* accession AtsD46 and AttD54, and synthetic allohexaploids TC54 and TC36. A_h, B_h, D_h, A_j, B_j, D_t correspond to homoeologs, where j subscript denotes Joyau, h for hexaploid and t for tauschii.

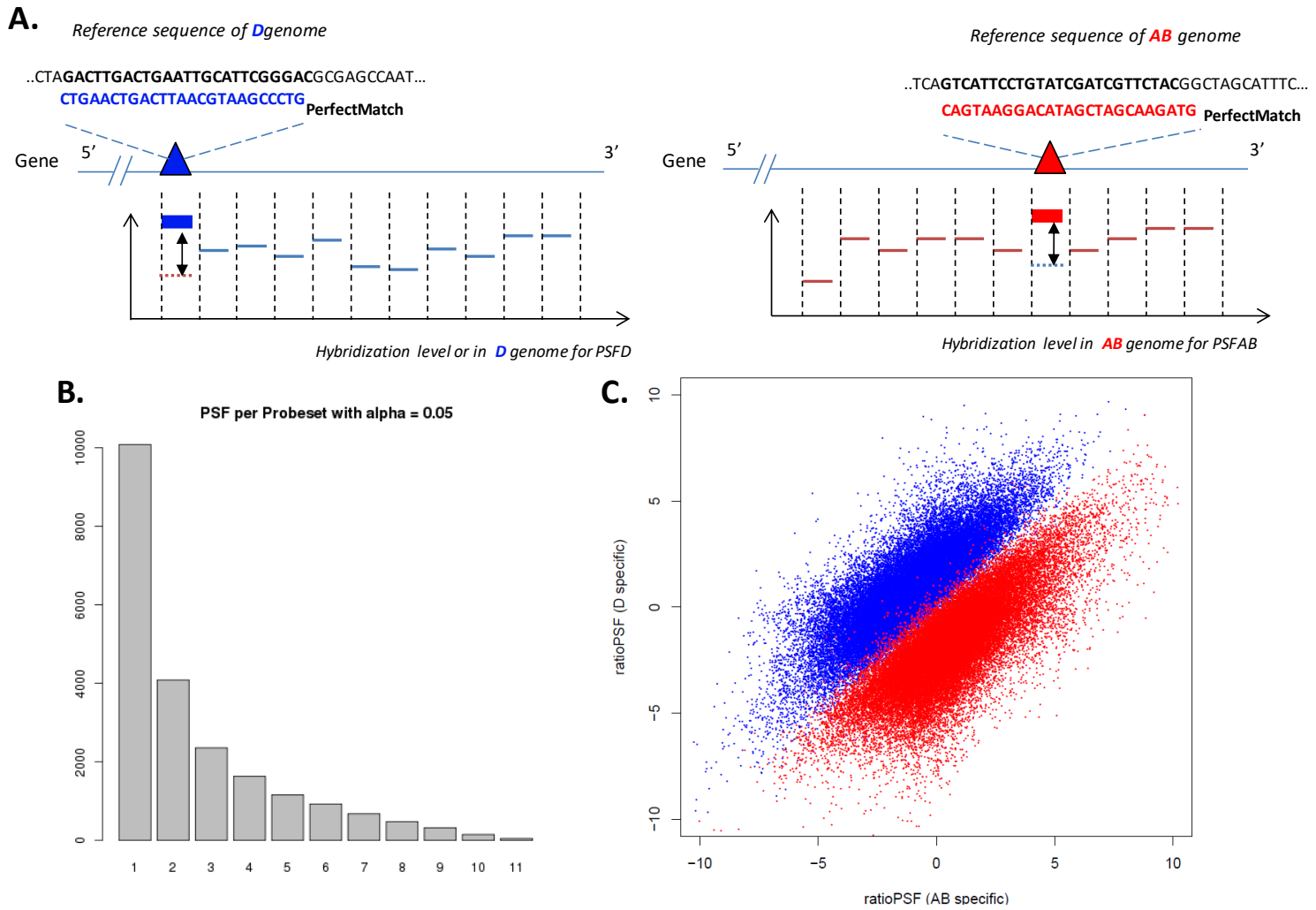


Figure 2: Schematic representation of Parent Specific feauture (PSF) discovery. **A.** Illustration of hybridization level of the 11 probes (PM) of a probeset, at 3' extremity of cDNA. PSFAB hybridize significantly more in BA genome and PSFD hybridizes significantly more in D genome. **B.** Frequency distribution of the number of PSFs per probeset. **C.** Scatterplot of $E_p/\bar{E}_t$ ratios in D (y-axis) and BA (x-axis) genome obtained for the 25 336 probes. E_p corresponds to PSF hybridization signal and $\bar{E}_t$ to the averaged hybridization signal of non-PSF(Akhunov et al.,2010).

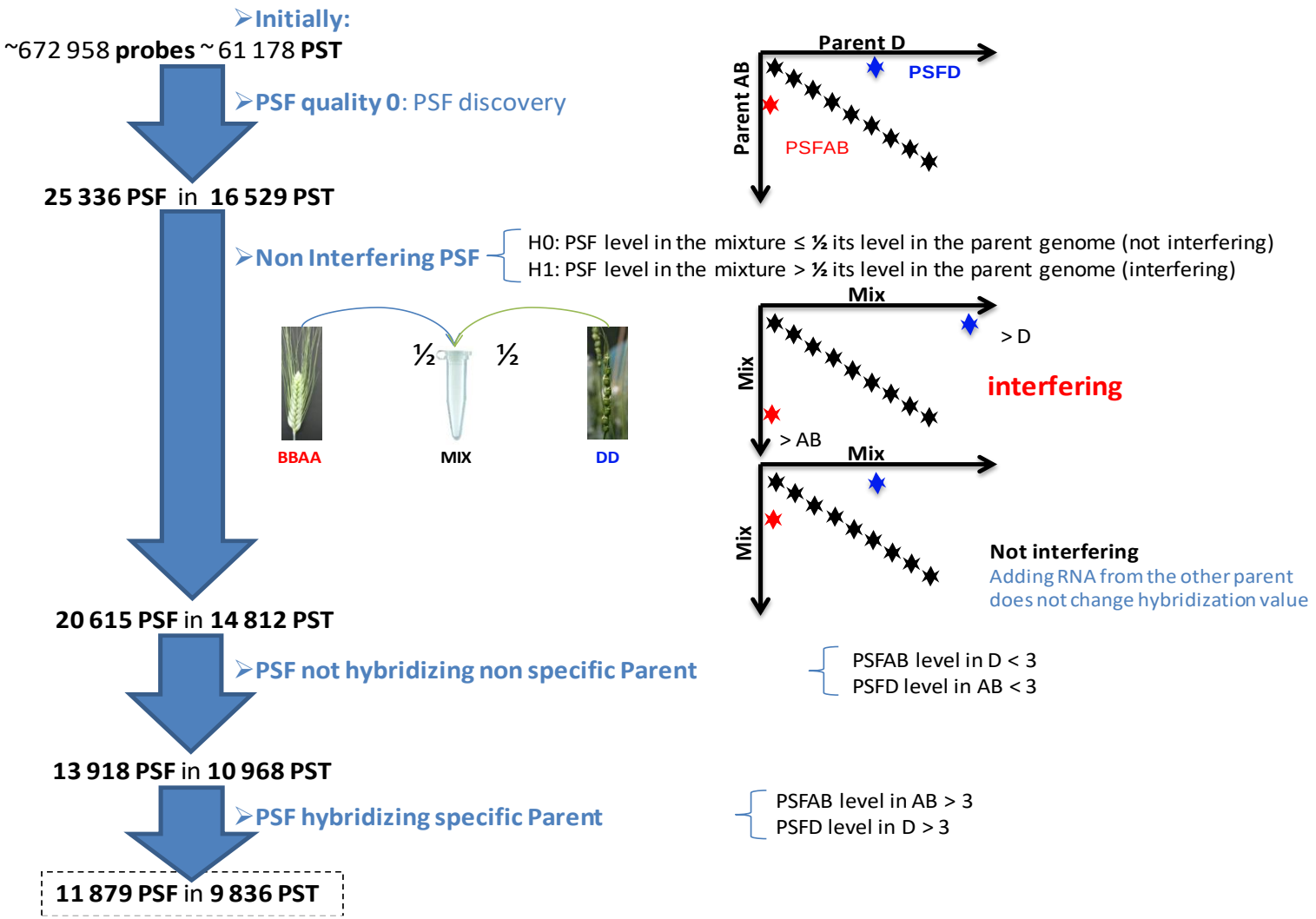


Figure 3: Parent specific features (PSF) discovery and assessing in relation to original probesets (PST). Starting with the 672 958 probes distributed in 61 178 PST, the first PSF discovery using limma test leaded to 25 336 PSF. Non interference test permits to remove those PSF, whose hybridization level in the MPV is significantly higher than ½ of the hybridization level in the specific parent.

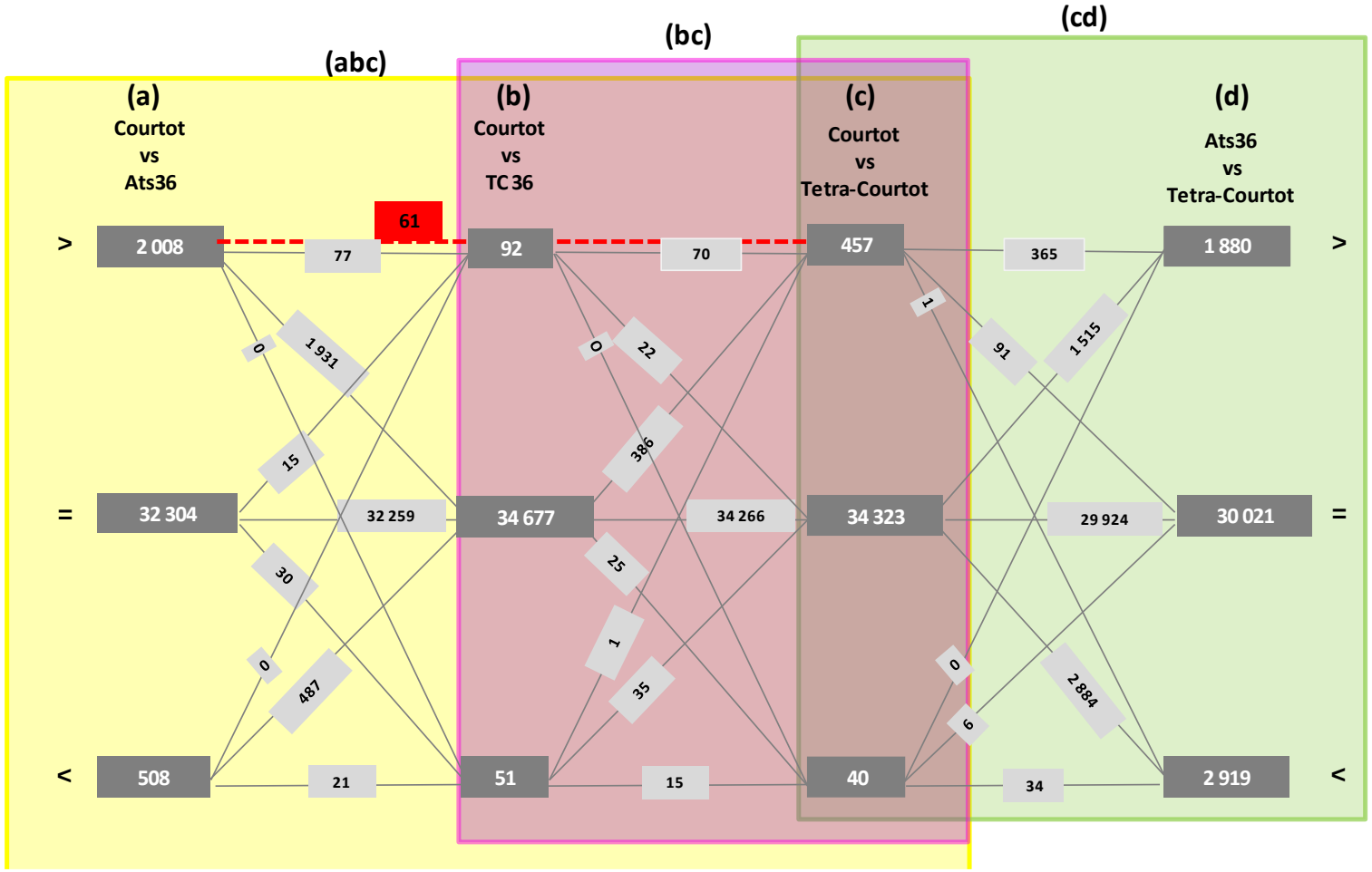


Figure 4: Cross comparison between categories of expression revealed by comparing various wheat species and genotypes.
(**bc**) Expression patterns revealed by comparing natural wheat allohexaploid cv Courtot and its extracted tetraploid "Tetra Courtot" matched with those revealed between Courtot and the newly synthetized allohexaploid TC36 and the natural allohexaploid cv Courtot; (**abc**) Expression patterns revealed by comparing the natural wheat allohexaploid cv Courtot and the synthetic wheat allohexaploid TC36; matched with those revealed between Courtot and the D genome donor *Ae. tauschii* (accession Ats36). The red dotted line detailed genes which are commonly up-expressed in the natural allohexaploid cv Courtot compared to TC36 and to Ats36. (**cd**) Expression patterns revealed by comparing the natural wheat allohexaploid cv Courtot and its extracted tetraploid "Tetra Courtot" ; matched with those revealed between the D genome donor *Ae. tauschii* (accession Ats36) and the extracted tetraploid "Tetra-Courtot". Numbers of genes shared by different expression categories are indicated on the cross-lines.

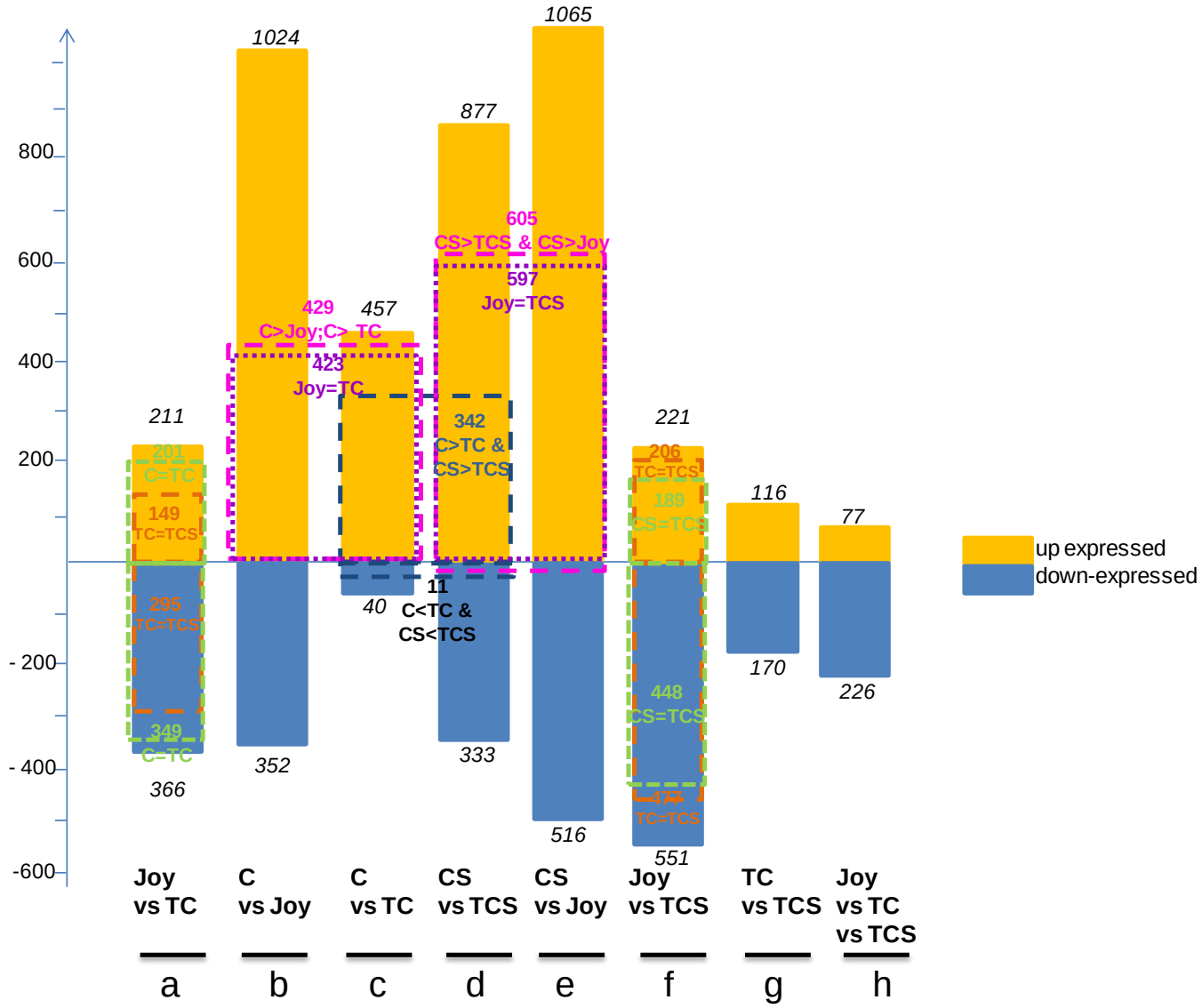


Figure 5: Summary of transcriptome comparisons differences between the natural allohexaploid cv Courtot (C) , Chinese Spring (CS) and their respective extracted allotetraploid "Tetra-Courtot" (TC) , "Tetra-Chinese Spring"(TCS) or the natural allotetraploid *T. turgidum* ssp *durum* cv Joyau (Joy); or between natural and extracted allotetraploids.

Histograms represent up- and down-expressed genes (relative to the first genotype mentioned in each comparison). Colored dotted parts correspond to amount of common genes for different comparisons, colored numbers correspond to their respective number of genes. Numbers in italic correspond to up- and down-expressed genes (orange and blue bars, respectively) for each comparison.

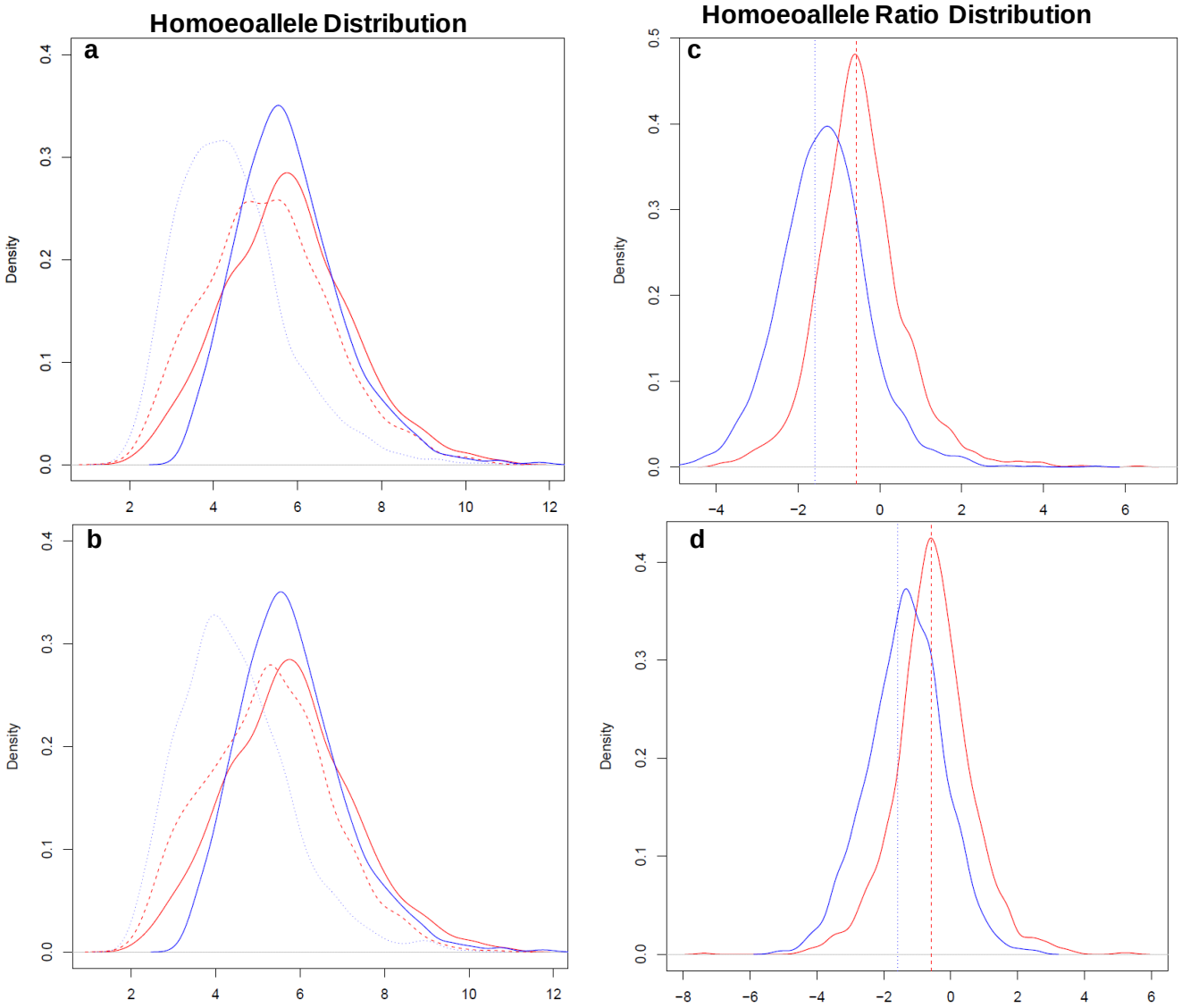


Figure 6: Density of B_hA_h, D_h and D_t subgenome expression in natural and newly-synthetized allohexaploids over extracted allotetraploids and *Ae.tauschii*.

a-b) Distribution of expression of B_hA_h, subgenome in extracted allotetraploids (plain red), in natural allohexaploids **(a**, dashed red lines) and in newly-synthetized allohexaploids **(b,** dashed red lines). Distribution of expression of D_t in the diploid species *Ae.tauschii* (blue plain line), of D_h in natural allohexaploids **(a**, dashed blue line) and in newly-synthetized allohexaploids **(b,** dashed blue lines).
c-d) Distribution of the ratio of B_hA_h subgenome expression level in natural allohexaploids over their expression in extracted allotetraploids (red); and distribution of D_h and D_t subgenome expression level in natural allohexaploids **(c)** and in the newly synthesized allohexaploids **(d)** over expression of D_t in diploid species (blue). The red dashed vertical line corresponds to the peak that equals to a ratio of 2/3 and the blue dashed vertical line to 1/3. Natural allohexaploids are represented by cv Coutot and Chinese Spring together, extracted allotetraploids by Tetra-Courtot and Tetra Chinese Spring, *Ae. tauschii* diploid species by accession D54 and D36, newly sytnhetized allohexaploids by TC36 and TC54.

Table 1: Details of plant material used in the analysis at the probeset level.
The table details species corresponding to the different genomes for expression comparison. For each species two biological replicates, represented by two different plants has been used, resulting in 22 Wheat Affymetrix arrays.

Genome	Ploidy level	Notation	Species and treatements
Extracted $B_hB_hA_hA_h$	tetraploid	B_hA_h	Tetra-Courtot (TC) Tetra-Chinese Spring (TCS)
Natural BBAA	tetraploid	B_jA_j	*T. Turgidum ssp. durum* cv Joyau (JOY)
Natural DD	diploid	D_t	*Ae. tauschii ssp. strangulata* (AtsD36) *Ae. tauschii ssp. tauschii* (AttD54)
Synthetic $B_hB_hA_hA_hD_tD_t$	allohexaploid	$B_hA_hD_t$	TC54 TC36
Natural $B_hB_hA_hA_hD_hD_h$	allohexaploid	$B_hA_hD_h$	*T. aestivum ssp. aestivum* cv Courtot (Courtot) *T. aestivum ssp. aestivum* cv Chinese Spring
Parental Mix $B_hB_hA_hA_hD_tD_t$	allohexaploid	$B_hA_hD_t$	$MPV^{36}TC$ $MPV^{54}TC$

Table 2: Parent specific features (PSF) and probeset with PSFAB and PSFD (PST) discovered and assessed. Filter1 corresponds to probes which pass limma test, and are named PSF. Filter2 corresponds to PSF which pass filter 1 and are not interfering, named non interfering PSF. Filter3 corresponds to non interfering PSF, which do not hybridize the non-specific Parent. Filter 4 selects non interfering PSF, which do not hybridize the non-specific Parent and hybridize in the specific Parent. PSF is for Parent Specific Features, PST for Probeset.

	Initial	Filter1	Filter2	Filter 3	Filter 4
PSF	**672 958**	**25 336**	**20615**	**13 918**	**11 879**
PSFAB	---	12 746	10 197	8 126	7197
PSFD	---	12 590	10 418	5 792	4682
PST	**61 178**	**16 529**	**14812**	**10 968**	**9 836**
PSTABD with PSFAB &PSFD satisfing conditions	---	4 605	2 956 PSFAB &PSFD non interfering	1190 PSFAB & PSFD non interfering, and not hybridizing non specific parent	514 PSFAB &PSFD non interfering, not hybridizing non specific parent, and hybridizing in the specific parent
PSTABD only PSFAB satisfing conditions	---		757 PSFAB non interfering	1942 PSFAB non interfering, and not hybridizing non specific parent	2286 PSFAB non interfering, not hybridizing non specific parent, and hybridizing in the specific parent
PSTABD only PSFD satisfing conditions	---		851 PSFD non interfering	985 PSFD non interfering, and not hybridizing non specific parent	1176 PSFD non interfering, not hybridizing non specific parent, and hybridizing in the specific parent
PSTAB	---	6 059	4 993	3 860	3 360
PSTD	---	5 865	5 255	2 991	2 500

Table 3: Cross comparisons between global gene level (GGE) in the extracted allotetraploid (B_hA_h) and natural allohexaploids ($B_hA_hD_h$), compared to PSFAB and PSFD expression level in natural allohexaploids. **a)** Direct cross comparisons; **b)** Cross comparisons using half of PSFAB expression level compared to total PSFD expression level in natural allohexaploids. For the allotetraploids B_hA_h, we consider Tetra-Courtot and Tetra-Chinese Spring together; for the allohexaploids $B_hA_hD_h$, we consider Courtot and Chinese Spring together. PSF: Parent Specific Features; GGE is the sum of PSFAB and PSFD.

a.

Expression of PSFAB vs PSFD in natural allohexaploid $B_hA_hD_h$

GGE in B_hA_h vs $B_hA_hD_h$	<	=	>	Total
<	44	7	2	53
=	18	224	218	460
>	0	0	1	1
Total	62	231	221	514

b.

Expression of ½ PSFAB vs PSFD in natural allohexaploid $B_hA_hD_h$

GGE in B_hA_h vs $B_hA_hD_h$	<	=	>	Total
<	48	3	2	53
=	44	284	134	460
>	0	0	1	1
Total	92	285	137	514

Table 4: Using PSF level analysis, cross comparisons between global gene gene expression (GGE) in extracted allotetraploids B_hA_h and natural allohexaploids $B_hA_hD_h$, compared to 2/3 PSFAB in extracted allotetraploid (B_hA_h) genome and PSFAB expression level in natural allohexaploids ($B_hA_hD_h$).

		2/3 PSFAB in extracted allotetraploids B_hA_h vs PSFAB in natural allohexaploids $B_hA_hD_h$			
		<	=	>	Total
GGE in B_hA_h vs $B_hA_hD_h$	<	0	53	0	53
	=	0	460	0	460
	>	0	1	0	1
	Total	0	514	0	514

Table 5: Using PSF level analysis, cross comparisons between global gene expression (GGE) in the diploid D_t genome donor *Ae.tauschii* and natural allohexaploids $B_hA_hD_h$, compared to 1/3 PSFD in *Ae.tauschii* as compared to PSFD expression level in the natural allohexaploids $B_hA_hD_h$.

		1/3 PSFD in diploid species D_t vs PSFD in natural allohexaploids $B_hA_hD_h$			
		<	=	>	Total
GGE in D_t vs $B_hA_hD_h$	<	0	115	8	123
	=	0	335	4	339
	>	3	49	0	52
	Total	3	499	12	514

Table 6: Using PSF level analysis, cross comparisons between global gene expression (GGE) in the extracted allotetraploid B_hA_h and diploids D_t genome donor *Ae. tauschii*; compared to PSFAB and PSFD expression level in the natural allohexaploid $B_hA_hD_h$. **a)** Direct cross comparisons**; b)** Cross comparisons using half of PSFAB expression level.

Expression of PSFAB vs PSFD in natural allohexaploids $B_hA_hD_h$

a

GGE in B_hA_h vs D_t	<	=	>	Total
<	58	73	0	131
=	4	148	93	245
>	0	10	128	138
Total	62	231	221	514

Expression of ½ PSFAB vs PSFD in natural allohexaploids $B_hA_hD_h$

b

GGE in B_hA_h vs D_t	<	=	>	Total
<	78	53	0	131
=	13	197	35	245
>	1	35	102	138
Total	92	285	137	514

Table 7: Cross comparisons using PSF level analysis, at global gene expression (GGE) between extracted allotetraploids (B_hA_h), natural allohexaploids ($B_hA_hD_h$) and the diploid *Ae. tauschiii* (D_t).

		GGE in B_hA_h vs GGE in $B_hA_hD_h$									
		<			=			>			
		GGE in $B_hA_hD_h$ vs D_t			GGE in $B_hA_hD_h$ vs D_t			GGE in $B_hA_hD_h$ vs D_t			Total
		<	=	>	<	=	>	<	=	>	
GGE in B_hA_h vs D_t	<	20	29	0	30	52	0	0	0	0	131
	=	0	1	3	2	224	15	0	0	0	245
	>	0	0	0	0	33	104	0	0	1	138
Total		20	30	3	32	309	119	0	0	1	514

Table 8: Using PSF level analysis, cross comparison between global gene expression (GGE) in extracted allotetraploids B_hA_h and newly synthetized allohexaploids $B_hA_hD_t$, compared to 2/3 PSFAB in AB genome and PSFAB expression level in $B_hA_hD_t$.

		2/3 PSFAB in extracted allotetraploids B_hA_h vs PSFAB in newly synthetized allohexaploids $B_hA_hD_t$			
		<	=	>	Total
GGE in B_hA_h vs in $B_hA_hD_t$	<	3	56	0	59
	=	0	443	0	443
	>	0	8	4	12
	Total	3	507	4	514

Table 9: Using PSF level analysis, cross comparison between global gene expression (GGE) in diploids D_t *Ae.tauschii* and newly synthetized allohexaploids $B_hA_hD_t$, compared to 1/3 PSFD in D_t genome and PSFD expression level in $B_hA_hD_t$.

		1/3 PSFD in diploid species D_t vs PSFD in newly synthetized allohexaploids $B_hA_hD_t$			
		<	=	>	Total
GGE in $D_t$vs in $B_hA_hD_t$	<	28	114	0	142
	=	8	294	5	307
	>	0	55	10	65
	Total	36	463	15	514

Table 10: Using PSF level analysis, between global gene level (GGE) in the extracted allotetraploid (B_hA_h) and diploid species *Ae. tauschii* (D_t) compared to PSFAB and PSFD expression level in newly-synthetized allohexaploids ($B_hA_hD_t$). **a)** Direct cross comparisons**; b)** Cross comparisons using half of PSFAB expression level.

Expression of PSFAB vs PSFD in newly synthetised allohexaploids $B_hA_hD_t$

a

GGE in B_hA_h vs D_t	<	=	>	Total
<	56	74	1	131
=	4	151	90	245
>	0	16	122	138
Total	60	241	213	514

Expression of ½ PSFAB vs PSFD in newly synthetised allohexaploids $B_hA_hD_t$

b

GGE in B_hA_h vs D_t	<	=	>	Total
<	89	42	0	131
=	22	187	36	245
>	0	36	102	138
Total	111	265	138	514

Table 11: Cross comparisons using PSF level analysis, between global gene expression (GGE) in the extracted allotetraploid (B_hA_h), diploid species *Ae.tauschii* (D_t), and newly-synthtetized allohexaploids wheat ($B_hA_hD_t$).

		GGE in B_hA_h vs GGE in $B_hA_hD_t$									
		<			=			>			
		GGE in $B_hA_hD_t$ vs D_t			GGE in $B_hA_hD_t$ vs D_t			GGE in $B_hA_hD_t$ vs D_t			Total
		<	=	>	<	=	>	<	=	>	
GGE in B_hA_h vs D_t	<	2	34	19	0	39	37	0	0	0	131
	=	2	2	0	28	204	7	0	0	2	245
	>	0	0	0	105	23	0	5	5	0	138
Total		4	36	19	133	266	44	5	5	2	514

Table S1: Overall number of expressed transcripts for genotypes and each replicates analyzed in the present study.

	Hybridization	Expressed I > 3		Not expressed I ⩽ 3	
Chinese Spring	CS2	28 271	26 946	33 019	31 251
	CS3	28 714		32 576	
Tetra Chinese Spring	tetraCS1	28 178	26 853	33 112	30 603
	tetraCS2	29 362		31 928	
Courtot	courtot_2	27 789	26 386	33 501	31 767
	courtotc_6j	28 120		33 170	
Tetra Courtot	tetra_courtotA_5b	28 724	26 660	32 566	31 039
	tetracourtot_2	28 187		33 103	
Joyau	JOY_1	28 620	27 342	32 670	29 648
	JOY_2	30 364		30 926	
TC54	SOSYN54_7	27 775	26 473	33 515	31 138
	SOSYN54_9	28 850		32 440	
TC36	S1SYN36_2	27 957	26 804	33 333	31 964
	Si_syn36B_7j	28 173		33 117	
MPV^{54}TC	mix_TC_54	28 114	27 255	33 176	32 049
	mix_TC_54bis	28 382		32 908	
MPV^{36}TC	mix_TC_36	27 975	27 057	33 315	32 407
	mix_TC_36bis	27 965		33 325	
AttD54	squa54a_6b	29 902	28 077	31 388	28 739
	squa54c_2j	30 726		30 564	
AtsD36	squa36A_3b	29 903	27 977	31 387	26 757
	squa36B_5j	32 607		28 683	

Table S2: Cross comparisons at global gene expression (GGE), using standard probeset level analysis, between allotetraploid natural cv Courtot (C), extracted « Tetra-Courtot » (TC), and the D genome donor diploid species *Ae. tauschii* accession AttD54. Values in smaller sizes correspond to detailled values when crossing with the two extracted allotetraploids.

	TC *vs* AttD54			
	>	=	<	
C > TC	1	111	345	457
C = TC	2 766	30 128	1 429	34 323
C < TC	37	3	0	40
Total	2 804	30 242	1 774	34 820

Table S3: Using PSF level analysis, cross comparison between global gene expression (GGE) in the newly-synthetized allohexaploids ($B_hA_hD_t$) and MPV, matched with the PSFD compared to half PSFAB expression levesl in the same newly-synthetized allohexaploids ($B_hA_hD_t$)

Expression of ½ PSFAB vs PSFD in newly synthetised allohexaploids $B_hA_hD_t$

GGE in MPV vs $B_hA_hD_t$	<	=	>	Total
<	3	9	21	33
=	102	243	112	457
>	6	13	5	24
Total	111	265	138	514

References

Adams, K.L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. J Hered 98, 136-141.

Adams, K.L., and Wendel, J. (2004). Exploring the genomic mysteries of polyploidy in cotton. Biological Journal of the Linnean Society 82, 573-581.

Adams, K.L., and Wendel, J.F. (2005). Novel patterns of gene expression in polyploid plants. Trends Genet 21, 539-543.

Adams, K.L., and Wendel, J.F. (2005b). Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8, 135-141.

Akhunova, A.R., Matniyazov, R.T., Liang, H., and Akhunov, E.D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics 11, 505.

Arnaud, D., Chelaifa, H., Jahier, J., and Chalhoub, B. (2013). Reprogramming of Gene Expression in the Genetically Stable Bread Allohexaploid Wheat. In Polyploid and hybrid genomics., Z. Chen and J.A. Birchler, eds (Ames : Wiley-Blackwell), pp. 195-211.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289-300.

Birchler, J.A., and Veitia, R.A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell 19, 395-402.

Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A. (2005). Dosage balance in gene regulation: biological implications. Trends Genet 21, 219-226.

Blake, N.K., Lehfeldt, B.R., Lavin, M., and Talbert, L.E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome 42, 351-360.

Chague, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., and Chalhoub, B. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol 187, 1181-1194.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans B, Corréa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VH, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CH, Wang X, Canaguier A, Chauveau A, Bérard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Wang X, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, and Wincker, P. (2014). Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. Science 345, 950-953

Chalupska, D., Lee, H.Y., Faris, J.D., Evrard, A., Chalhoub, B., Haselkorn, R., and Gornicki, P. (2008). Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci U S A 105, 9691-9696.

Chelaifa, H., Chague, V., Chalabi, S., Mestiri, I., Arnaud, D., Deffains, D., Lu, Y., Belcram, H., Huteau, V., Chiquet, J., Coriton, O., Just, J., Jahier, J., and Chalhoub, B. (2013).

Prevalence of gene expression additivity in genetically stable wheat allohexaploids. New Phytol 197, 730-736.
Chen, Z.J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58, 377-406.
Chen, Z.J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. Nat Rev Genet. 14, 471-482.
Comai, L. (2005). The advantages and disadvantages of being polyploid. Nat Rev Genet. 6, 836-846.
Delmar, P., Robin, S., and Daudin, J.J. (2005). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics 21, 502-508.
Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S., and Wendel, J.F. (2008). Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42, 443-461.
Dvorak, J., and Zhang, H.B. (1990). Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proc Natl Acad Sci U S A 87, 9640-9644.
Dvorak, J., Terlizzi, P., Zhang, H.B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome 36, 21-31.
Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. BMC Biol 6, 16.
Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., and Osborn, T.C. (2007). Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19, 3403-3417.
Griffiths, S., Sharp, R., Foote, T., Bertin, I., Wanous, M., Reader, S., Colas, I., and Moore, G. (2006). Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature 439, 749-752.
Grover, C.E., Yu, Y., Wing, R.A., Paterson, A.H., and Wendel, J.F. (2008). A phylogenetic analysis of indel dynamics in the cotton genus. Mol Biol Evol 25, 1415-1428.
Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homeolog expression bias and expression level dominance in allopolyploids. The New Phytologist. 196, 966-971.
Guan, X., Pang, M., Nah, G., Shi, X., Ye, W., Stelly, D.M., and Chen, Z.J. (2014). miR828 and miR858 regulate homoeologous MYB2 gene functions in Arabidopsis trichome and cotton fibre development. Nat Commun 5, 3050.
Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K., Chapman, E., Carrington, J., Chen, X., Wang, X., and Chen, Z. (2009). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci U S A. 106, 17835-17840.
Hegarty, M.J., Jones, J.M., Wilson, I.D., Barker, G.L., Coghill, J.A., Sanchez-Baracaldo, P., Liu, G., Buggs, R.J., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2005). Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. Mol Ecol 14, 2493-2510.
Hovav, R., Udall, J.A., Chaudhary, B., Rapp, R., Flagel, L., and Wendel, J.F. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc Natl Acad Sci U S A 105, 6191-6195.
Hu, Z., Han, Z., Song, N., Chai, L., Yao, Y., Peng, H., Ni, Z., and Sun, Q. (2013). Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (Triticum aestivum). New Phytol 197, 1344-1352.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002a). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc Natl Acad Sci U S A 99, 8133-8138.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002b). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proceedings of the National Academy of Sciences, USA 99, 8133-8138.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS One. 5, e12336.

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

Kerber, E.R. (1964). Wheat: Reconstitution of the Tetraploid Component (AABB) of Hexaploids. Science 143, 253-255.

Kihara, H. (1944). Discovery of the DD-analyser, one of the ancestors of vulgare wheats. Ag. Hort. (Tokyo) 19, 889-890.

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

Liu, B., Xu, C., Zhao, N., Qi, B., Kimatu, J.N., Pang, J., and Han, F. (2009). Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. J Genet Genomics 36, 519-528.

Madlung, A., and Wendel, J.F. (2013). Genetic and epigenetic aspects of polyploid evolution in plants. Cytogenet Genome Res 140, 270-285.

McFadden, E.S., and Sears, E.R. (1946). The origin of Triticum speltoides and its free-threshing hexaploid relatives. J. Hered. 37, 81-89.

Mestiri, I., Chague, V., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Chalhoub, B., and Jahier, J. (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. New Phytol 186, 86-101.

Meyers, L.A., and Levin, D.A. (2006). On the abundance of polyploids in flowering plants. Evolution 60, 1198-1206.

Nesbitt, M., and Samuel, D. (1996). From the staple crop to extinction? The archaeology and history of hulled wheats. . . In Hulled Wheats. Proceedings of the First International Workshop on Hulled Wheats. (International Plant Genetic Resources Institute, Rome, Italy.).

Nomura, T., Ishihara, A., Yanagita, R.C., Endo, T.R., and Iwamura, H. (2005). Three genomes differentially contribute to the biosynthesis of benzoxazinones in hexaploid wheat. Proc Natl Acad Sci U S A 102, 16490-16495.

Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., and Martienssen, R.A. (2003). Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19, 141-147.

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

Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M., and Chen, Z.J. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome Biol 10, R122.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A 101, 9903-9908.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.J., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.H., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., ur Rahman, M., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.K., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A., Zhang, T., Dennis, E.S., Mayer, K.F., Peterson, D.G., Rokhsar, D.S., Wang, X., and Schmutz, J. (2011). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492, 423-427.

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Dolezel, J., Fahima, T., Budak, H., Keller, B., Salvi, S., Maccaferri, M., Steinbach, D., Feuillet, C., Quesneville, H., and Salse, J. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J 76, 1030-1044.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3, 827-837.

Pumphrey, M., Bai, J., Laudencia-Chingcuanco, D., Anderson, O., and Gill, B.S. (2009). Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics 181, 1147-1157.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F., and Liu, B. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. BMC Biol 10, 3.

Rambani, A., Page, J.T., and Udall, J.A. (2014). Polyploidy and the petal transcriptome of Gossypium. BMC Plant Biol 14, 3.

Rapp, R.A., and Wendel, J.F. (2005). Epigenetics and plant evolution. New Phytol 168, 81-91.

Rapp, R.A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. BMC Biol 7, 18.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol Evol 16, 351-358.

Riley, R., and Chapman, V. (1958). Genetic control of the cytologically diploid behaviour of hexaploid wheat. Nature 182, 713-715.

Riley, R., Unrau, J., and al., e. (1958). Evidence on the origin of the B genome of wheat. J. Hered. 49, 91-98.

Rostoks, N., Mudie, S., Cardle, L., Russell, J., Ramsay, L., Booth, A., Svensson, J.T., Wanamaker, S.I., Walia, H., Rodriguez, E.M., Hedley, P.E., Liu, H., Morris, J., Close, T.J., Marshall, D.F., and Waugh, R. (2005). Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274, 515-527.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and 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.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments.
. Statistical Applications in Genetics and Molecular Biology 3.
Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu Rev Plant Biol 60, 561-588.
Song, K., Lu, P., Tang, K., and Osborn, T.C. (1995). Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc Natl Acad Sci U S A 92, 7719-7723.
Stamati, K., Mackay, I., and Powell, W. (2009). A quantitative genomic imbalance gene expression assay in a hexaploid species: wheat (Triticum aestivum). Genome 52, 89-94.
Takumi, S., Nasuda, S., and al., e. (1993). Wheat phylogeny determined by RFLP analysis of nuclear DNA. I. Einkorn wheat. Jpn. J. Genet. 68, 73-79.
Talbert, L.E., Blake, N.K., Storlie, E.W., and Lavin, M. (1995). Variability in wheat based on low-copy DNA sequence comparisons. Genome 38, 951-957.
Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science 320, 486-488.
Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat Rev Genet 10, 725-732.
Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., and Chen, Z.J. (2006). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics 172, 507-517.
Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol 42, 225-249.
Xu, C., Bai, Y., Lin, X., Zhao, N., Hu, L., Gong, Z., Wendel, J.F., and Liu, B. (2014). Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. Mol Biol Evol 31, 1066-1076.
Yang, T., Furuta, Y., Nagata, S., and Watanabe, N. (1999). Tetra Chinese Spring with AABB genomes extracted from the hexaploid common wheat, Chinese Spring. . Genes & Genetic Systems 74, 67-70.
Yoo, M.J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) 110, 171-180.
Zhang, H., Bian, Y., Gou, X., Zhu, B., Xu, C., Qi, B., Li, N., Rustgi, S., Zhou, H., Han, F., Jiang, J., von Wettstein, D., and Liu, B. (2013). Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. Proc Natl Acad Sci U S A 110, 3447-3452.
Zhang, H., Zhu, B., Qi, B., Gou, X., Dong, Y., Xu, C., Zhang, B., Huang, W., Liu, C., Wang, X., Yang, C., Zhou, H., Kashkush, K., Feldman, M., Wendel, J.F., and Liu, B. (2014). Evolution of the BBAA Component of Bread Wheat during Its History at the Allohexaploid Level. Plant Cell.
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., and Chalhoub, B. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci U S A 108, 18737-18742.

3.3. Analyses et Discussion complémentaires

Les deux méthodes d'analyse de l'expression globale du gène, présentées dans cet article, sont :

- la méthode standard à l'échelle du probeset, qui fait la moyenne de l'ensemble des 11 oligos-sondes contenues dans le probeset. Ceci a permis de comparer l'expression de 34 820 gènes;

- et la méthode originale à l'échelle de la sonde, dans laquelle l'expression globale du gène, dans un génome allohexaploïde, est la somme des hybridations des sondes PSFAB et PSFD. Ceci a permis de comparer l'expression de 514 gènes.

Ainsi, 374 gènes sont conjointement analysés par les deux méthodes. La comparaison pour ces gènes de l'expression globale estimée par les deux méthodes montre que 368/374, soit 98.4% des gènes en communs, sont à un niveau d'expression égal entre l'allohexaploïde naturel et l'allohexaploïde synthétique (Fig.40). Seulement 6/374~1.6% gènes restant sont sur- ou sous-exprimés dans un des allohexaploïdes avec la méthode « PSF » tandis qu'ils sont à expression globale égale avec la méthode probeset. Ceci confirme l'adéquation de la méthode PSF tel que utilisée ici.

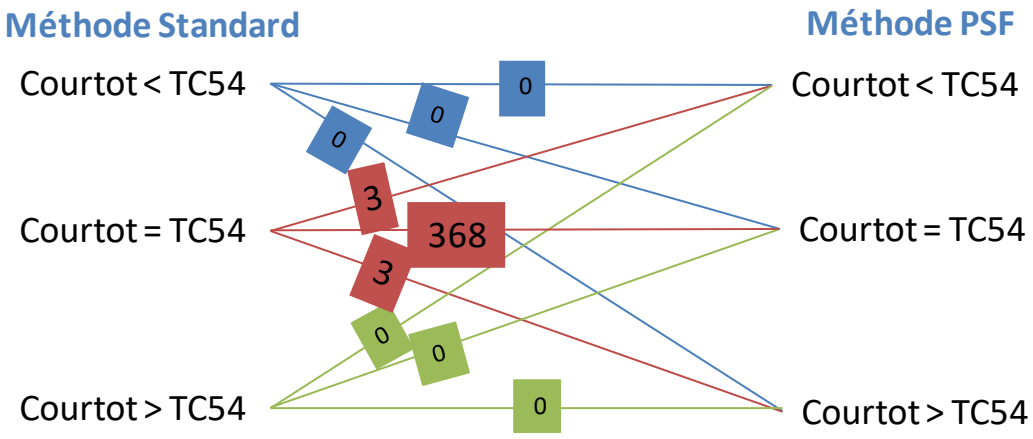


Figure 40: Comparaisons croisées entre la méthode probeset et la méthode PSF pour analyser estimer l'expression globale des gènes du blé. Sont comparés ici le blé hexaploïde naturel cv Courtot et le blé hexaploïde synthétique TC54. Les chiffres dans les boites reliant les comparaisons des deux méthodes représentent les 374 gènes communs.

Cette étude d'analyse de l'expression des sous-génomes B_hA_h, D_h ou D_t lors du changement du niveau de ploïdie, révèle une diminution de l'expression de ces sous-génomes dans le blé allohexapoïde par rapport à l'allotétraploïde ou au diploïde, suggérant que leurs niveaux d'expression sont inversement proportionnels au niveau de ploïdie.
Nous avons pu préciser la contribution de ces sous-génomes à l'expression globale dans les allohexaploïdes naturels et synthétiques. La majorité des gènes analysés dans l'allohexaploïde ont une expression globale contribuée par 2/3 par le sous-génome B_hA_h et par 1/3 de l'expression par le sous-génome D_t du diploïde. Par ailleurs, dans l'allohexaploïde, l'expression du sous-génome B_hA_h est souvent égale au 2/3 de son expression dans l'allotétraploïde et celle du sous-génome D est égale au 1/3 de son expression dans l'espèce diploïde.

Toutefois, dans cette étude nous n'avons pas pu disséquer l'expression des homéologues A_h et B_h de façon indépendante. Aussi l'outil microarray reste moins précis que les nouvelles technologies de séquençage, type RNA-Seq, qui permettent une définition précise des homéologues et de leur niveau d'expression et que je développe dans le chapitre qui suit.

Chapitre 4

Dissection de l'expression des homéologues dans les blés allopolyploïdes

4.1. Contexte et questions posées

Le développement des NGS a constitué un essor considérable de l'analyse haut-débit des génomes, et notamment des génomes complexes présentant des niveaux de ploïdie élevés. Ces outils de séquençage haut-débit permettraient par leur résolution de mieux comprendre la composition des génomes polyploïdes et le devenir des gènes dupliqués.

Le séquençage 'brouillon' du génome du blé hexaploïde, très récent (IWGSC, 2014), a permis de montrer qu'il présente une très grande conservation des séquences codantes entre les sous-génomes A_h, B_h et D_h (97% d'identité) et ceux de ses progéniteurs (IWGSC, 2014). Ce séquençage a rendu disponibles les séquences des homéologues A_h, B_h et D_h (triplets) (IWGSC, 2014) de 8605 gènes. Ceci m'a permis une première approche de disséquer l'expression des gènes en celle de leurs homéologues, utilisant le séquençage haut-débit des ARNm (RNA-Seq) et sa cartographie 'unique' sur ces séquences homéologues.

4.1.1. Les données et leur prétraitement

Les technologies de séquences NGS diffèrent sur la taille des lectures générées (allant de 35pb à 100Kpb), leur rendement, la durée des réactions (ou 'run'), le coût du séquenceur et des réactifs, le type d'erreur et le taux d'erreur (Henson et al., 2012 ; Liu et al., 2012b). Suivant la problématique et le modèle biologique utilisé, le choix de la technologie est important. Dans ce projet, nous avons choisi Illumina HiSeq 2000, pour le séquençage des mRNA préparés à partir des feuilles, dit RNA-Seq, en raison du coût faible des réactifs et du faible taux d'erreurs de séquences (Henson et al., 2012).

Une fois les séquences disponibles, elles ont été alignées sur les 8605 triplets du génome du blé hexaploïde par l'outil BWA. Afin d'effectuer un alignement spécifique et de distinguer l'expression des homéologues, j'ai sélectionné les lectures en alignement unique, c'est-à-dire un alignement avec une tolérance zéro (0 Mismatch).

4.1.2. Le traitement des données

Dans mes travaux, j'ai utilisé la normalisation RPKM (Reads Per Kilobase of exon per Million mapped reads) (Mortazavi et al., 2008) car je possède des données de séquences RNA-Seq en lecture simple (une seule extrémité est séquencée), qui est très utilisée (Rambani

et al., 2014) tout comme son analogue FPKM utilisé pour les RNA-Seq avec un séquençage aux deux extrémités (ou lectures 'pairées') (Choulet et al., 2014; IWGSC, 2014 ; Pfeifer et al., 2014). Cette normalisation permet de normaliser les comptages bruts des lectures RNA-Seq alignées sur les séquences des gènes de références (ici les 8605 triplets de gènes), par rapport à 1kb de la longueur des gènes et 1 million de lectures alignées sur les gènes de références.

Au début des analyses, une interrogation s'est posée sur l'adéquation de la normalisation RPKM pour comparer l'expression de chaque homéologue entre les espèces de trois niveaux de ploïdie. En effet, la normalisation des lectures par million de lectures alignées, ne permettraient pas de réaliser les comparaisons d'expression si les proportions de lectures alignées étaient différentes entre les différents niveaux de ploidie. Les comparaisons (Table S1 de l'article) montre une proportion similaire de ~10% de lectures alignées. Il n'a donc pas été nécessaire de trouver une autre normalisation reflétant le nombre de lectures alignées.

L'analyse différentielle permettant d'identifier les homéologues ou les triplets de gènes différentiellement exprimés utilise DESEQ (Anders and Huber, 2010), qui est une méthode paramétrique bien adaptée aux échantillons de petite taille (ici, dans nos travaux, n=3) (Sims et al., 2014).

DESeq utilise un modèle basé sur la distribution binomiale négative car les données RNASeq sont discrètes et sur-dispersées. En effet, la variance est très supérieure à la moyenne: il y a une sur-dispersion des données. Les données sont donc approchées par la loi Binomiale négative, avec K_{ij} la variable aléatoire du nombre de lectures dans l'échantillon j, assignées au gène i:

$$K_{ij} \sim BN(\mu_{ij} ; \sigma^2_{ij}), \text{ avec } \mu_{ij} \text{ et } \sigma^2_{ij} \text{ non nécessairement égaux.}$$

μ_{ij} et σ^2_{ij} sont inconnues et σ^2_{ij} peut s'écrire : $\sigma^2_{ij} = \mu_{ij}(1+\phi_{ij}\mu_{ij})$, où ϕ_{ij} représente la dispersion.

Estimer la variance σ^2_{ij} et la moyenne μ_{ij} revient à estimer la dispersion ϕ_{ij} et la moyenne μ_{ij}.

DESeq permet une estimation de la dispersion (Anders and Huber, 2010) et estime ϕ_{ia} et μ_{ia} pour chaque condition a (ici chaque génotype a). La moyenne estimée $\mu_{ia\ estimée}$ et la variance estimée $\sigma^2_{ia\ estimée}$ empiriquement pour chaque génotype et chaque gène permettent de déduire la dispersion empirique pour chaque génotype et chaque gène.

Pour chaque gène i et chaque génotype a, DESeq considère deux estimateurs possibles de la dispersion :

− $\phi_{ia\ estimée,}$ qui est dérivée des estimations de $\mu_{ia\ estimée}$ et $\sigma^2_{ia\ estimée}$;

– $\tilde{\phi}_{ia}$, qui utilise les données sur tous les gènes avec l'hypothèse que $\tilde{\phi}_{ia} = f(\mu_{ia\ estimée})$. Cet estimateur calcule les estimations de $\mu_{ia\ estimée}$ et $\sigma^2_{ia\ estimée}$ pour tous les gènes et ajuste une régression f entre $\mu_{a\ estimée}=(\mu_{ia\ estimée})_{i=1\ à\ n\ gènes}$ et $\sigma^2_{a\ estimée}= (\sigma^2_{ia\ estimée})_{i=1\ à\ n\ gènes}$.

DESeq choisit alors la plus grande valeur entre $\phi_{ia\ estimée}$ et $\tilde{\phi}_{ia}$.
Puis DESeq teste la différence d'expression entre les deux conditions étudiées, en utilisant la loi binomiale négative pour laquelle les paramètres μ_{ia} et ϕ_{ia} ont été ainsi estimées.

4.2. Partionnement de l’expression des gènes homéologues en diminuant puis ré-augmentant la polyploïdie du blé

Ce chapitre est présenté sous forme d’article, en vue d’une soumission.

Duplicate and partitioning of homoeolog gene expression in the decreasing and re-increasing wheat polyploid model

Smahane Chalabi[1], Chelaifa Houda[1], LeFloch Edith[1,2], Karine Labadie[3], Matthias Pfeifer[4], Dominique Arnaud[1], Mestiri Imen[1], Vinh Ha DinhThi[1], Isabelle LeClainche[1], Corinne DaSilva[3], Denise Deffains[5], Virginie Huteau[5], Olivier Coriton[5], Claudine Devauchelle[2], Carene Rizzon[2], Joseph Jahier[5] Julien Chiquet[2], Harry Belcram[1], Eduard Akhunov[6], Patrick Wincker[3], Klaus Mayer[4], and Boulos Chalhoub[1].

[1]Unité de Recherche en Génomique Végétale URGV (INRA-CNRS – UEVE), Organization and Evolution of Plant Genomes, 91057,Evry Cedex, France;
[2]Laboratoire Statistique et Génome, Université d'Evry Val d'Essonne, UMR CNRS 8071 - USC INRA, Evry, France
[3]Commissariat à l'Energie Atomique (CEA), Genoscope, 91057,Evry Cedex, France;
[4]Plant Genome and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.
[5]Unité Mixte de Recherches INRA, Agrocampus Rennes - Université Rennes 1, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), 35653, Le Rheu, France ;
[6]Department of Plant Pathology, Throckmorton Plant Sciences Center, Kansas State University, Manhattan, KS 66506, USA.

[*]Corresponding author email: chalhoub@evry.inra.fr

Abstract

The reprogramming of gene expression may represent an adaptive mechanism for stable polyploid species. The present study aims to unveil global gene expression changes and its partitioning between constituent homoeologs across the different ploidy levels of wheat. This was achieved by the availability of 8605 triplets of genes with A_h, B_h and D_h homoeologs sequences from hexaploid wheat that have been used dissect homoeolog expression in an original wheat material, with decreasing and re-increasing ploidy levels, using massive parallel mRNA sequencing.

Our results indicate the majority of homoeologs contribute to gene expression, the prevalence of genome equivalence and the absence of significant bias towards any subgenome, in the recent wheat allohexaploid, but also in the older wheat allotetraploid. A main finding is that the majority of homoeologs generally increase expression when separated and decrease expression when joined again together. This is demonstrated by the fact that expression of A_h and B_h homoeologs in natural and resynthesized allohexaploid wheat are generally 2/3 of their expression in the extracted allotetraploid 'Tetra-Courtot', whereas those of the D_h and D_t homoeologs in allohexaploids are 1/3 of that of the D_t genes in the diploid *Ae. tauschii*.

Results obtained in this work contribute to our understanding of gene expression regulation at different ploidy levels by dissecting the global gene expression at the level of constituent homoeologs. The future functional analysis of the different gene expression categories would reveal important gene functional categories that are regulated in response to polyploidy.

Introduction

Polyploidy is a fundamental driver of biodiversity with significant consequences on genome structure, organization and evolution (Soltis and Soltis, 2012). It plays a major role in angiosperm evolution, plant speciation and diversification (Doyle et al., 2008; Leitch and Leitch, 2008; Soltis and Soltis, 2009).

Following polyploidization, duplicated genes follow one of many possible evolutionary fates such as non-functionalization consisting in gene deletion or pseudogenization. Duplicated genes can also be retained as functionally redundant, neo-functionalized (evolution of novel functions among alleles or homoeoalleles) (Zhang et al., 2011) or sub-functionalized (evolution of partitioned ancestral functions among alleles or homoeoalleles) (Prince and Pickett, 2002; Zhang et al., 2011), providing the potential for novelty and plasticity of polyploid genomes (Comai, 2005 ; Chen, 2007; Leitch and Leitch, 2008; Soltis and Soltis, 2009; Pont et al., 2013). These are ongoing processes during the evolution of polyploid genomes (Chen, 2007; Gaeta et al., 2007).

Wheat species provide a good example of relatively recent and stable allopolyploids. The widely-cultivated allohexaploid wheat *T. aestivum* L. (2*n*=6*x*=42, BBAADD), also known as common or bread wheat, originated as the result of two separate amphiploidization events. The allotetraploid *T. turgidum* L. (2*n*=4*x*=28, BBAA) arose less than 0.5 million years ago as a result of hybridization between *T. urartu* Tumanian ex Gandylian (2*n*=2*x*=14, AA) and an unidentified diploid *Aegilops* species of the section *Sitopsis,* thought to be *Ae. speltoides* (2*n*=2*x*=14, SS) or a close relative thereof, as the donor of the B genome (Riley et al., 1958 ; Dvorak and Zhang, 1990 ; Dvorak et al., 1993 ; Takumi et al., 1993 ; Talbert et al., 1995; Blake et al., 1999; Huang et al., 2002b ; Chalupska et al., 2008). A spontaneous hybridization between the early-domesticated tetraploid *T. turgidum* ssp. *dicoccum* and the diploid goatgrass *Ae. tauschii* (2*n*=2*x*=14, DD), about 10 000 years ago, gave rise to *T. aestivum* (Kihara, 1944 ; McFadden and Sears, 1946 ; Nesbitt and Samuel, 1996; Huang et al., 2002b).

Allopolyploidy makes homoeolog expression studies difficult in that the high sequence similarity among duplicated genes did not facilitate the development of a microarray homoeolog-specific approach, where most attempts remains with a relatively limited number of genes and didn't prevent cross hybridizations (Udall et al., 2006; Akhunova et al., 2010; Flagel and Wendel, 2010). Therefore, most studies in polyploids, including wheat, have characterized effects of alloploidy on global gene expression (not

separating the constitutent homoeologs) by comparing natural or synthetic allopolyploids to their progenitors or average of their progenitors (Adams and Wendel, 2004 ; Hegarty et al., 2005; Wang et al., 2006b; Pumphrey et al., 2009; Rapp et al., 2009; Akhunova et al., 2010; Chague et al., 2010; Chelaifa et al., 2013 ; Zhang et al., 2014).

Massive mRNA sequencing and comparison become an alternative to microarrays for transcriptome profiling, and precisely for characterizing homoeolog-specific expressions in polyploid organisms (Chalhoub et al., 2014). The rise of next-generation sequencing (NGS) allows an accuracy of homoeologs identification for hybrids and polyploids, although only few studies have been done at homoeologous gene expression levels (Higgins et al., 2012 ; Bell et al., 2013 ; Page et al., 2013b ; Roulin et al., 2013 ; Yoo et al., 2013 ; Chalhoub et al., 2014).

The recent draft genome sequencing of the ahexaploid bread wheat genome (IWGSC, 2014) reveals 124,201 genes, with about 35 000 genes for each of the A, B and D subgenomes. Among these, 8605 are triplets of genes with three homoeologs, belonging to A, B and D subgenomes.

The present study exploit the availability of these triplet sequences and the possibility of extracting allotetraploid genome $B_hB_hA_hA_h$ from natural allohexaploid wheat cv Courtot, as well as resynthesizing a highly stable allohexaploid wheat (TC109) (Mestiri et al., 2010), in order to unveil global gene expression changes and its partitioning and regulation between the different constituent homoeologs across the different ploidy levels.

Materials and methods

Plant material

Five genotypes have been used in this study (Figure 1). *Ae. tauschii* ssp. *strungulata* acc. 109 (D_t, 2n=2x=14) (the subscript t denotes *tauschii*) was shown to give the most stable synthetic wheat allohexaploids in term of stability of homologous paring and a very low aneuploid frequency (Mestiri et al., 2010). The natural allohexaploid *T. aestivum* ($B_hA_hD_h$, 2n=6x=42) (the subscript h denotes hexaploidy) is the French cultivar Courtot. "Tetra-Courtot" is the tetraploid (B_hA_h, 2n=4x=28) extracted from cv Courtot (by six successive backcrosses) (Mestiri et al., 2010). The natural allotetraploid *T.turgidum* used here is the French cv Joyau (B_jA_j, 2n=4x=28) (the j subscript denotes Joyau). TC109 is the most stable resynthesized hexaploid wheat that we have previously characterized (Mestiri et al., 2010), having the extracted tetraploid Tetra-Courtot as B_hA_h genome donor and *Ae. tauschii* spp. strungulata acc.109 (D109) as the D genome donor (Mestiri et al., 2010). We chose in this study the second-selfed (S1) generation of this resynthetized allohexaploid wheat.

All plants of these genotypes were grown in growth chambers at 22°C and 16h day length. Three different plants, representing three biological replicates, were analyzed per each genotype.

RNA extractions, RNA-Seq libraries and sequencing

RNA samples were extracted from whole leaves of the three replicates according to (Chague et al., 2010 ; Chelaifa et al., 2013; Chalhoub et al., 2014), then samples were quantified and evaluated for their quality on an Agilent Bioanalyzer. The 15µg leaves cDNA libraries were realized by TruSeq RNA sample preparation kits (Illumina), which allows to convert mRNA sample into a library of suitable matrices molecules for cluster generation. cDNA libraries were indexed per genotype from 2 to 12, and sequenced twice on 7 lanes by multiplexing (6 librairies/lane) of Illumina Solexa Hiseq2000 flow-cell (100 bp single-end read run).

Analysis of homoeologous gene expression

We used the Illumina single-end mRNA-Seq reads to measure transcript abundance in leaf tissues of wheat plants, using three biological replicates with an average of ~50 million single-end reads per replicate (details in Table S1).

Homoeologous gene triplets

We used the 8605 genes where the sequences of the three homoeologs, belonging to A_h, B_h and D_h subgenomes (triplets) were rendered available recently (IWGSC, 2014) (Figure S1) to evaluate homoeologous gene expression changes at the different ploidy levels compared here. The average mean length of these genes is 1342 bp, and median is 1137 bb (with a length of genes ranging from 294 to 14230 bp) with an average 90% overlapping sequences between the three homoeologs.

Mapping and counting of mRNA sequence reads

The measurement and discrimination of expression was rendered possible through unique mapping of mRNA-Seq reads based on sequence differences between A_h, B_h and D_h homoeologs (Chalhoub et al., 2014). mRNA-Seq reads were mapped using BWA (Li and Durbin, 2010) with default parameters where 0 mismatch and minimum reads length of 35bp were applied (Version: 0.6.1-r104, seed 35, gap penalty 11). Mapped RNA-Seq reads were then filtered using SAMtools (Version: 0.1.12a) and only unique matches were considered (Table S1).

Homoeologous genes were considered to be expressed and included in statistical analysis, if they show one or more mapped reads, leading thus to 8528 triplets that have been considered here.

Data and read count normalization

Since we aimed at comparing expression of each homoeolog in different genotypes having different ploidy levels, we used the standard RPKM (Reads Per Kilobase per Million mapped reads) normalization (Mortazavi et al., 2008), which normalize both in terms of the library read depth and the homoeolog length.

Statistical Analysis

After the normalization step, a principal component analysis (PCA) was applied to quickly summarize the data and look for spurious technical effects. Figure S2 shows the projection of the 15samples on the first two PC-axes in the sample space, with satisfactory reproducibility between biological replicates, explaining almost 88.06 % of the variance.

Comparison of global and homoeolog expression

The differential analysis was performed, on the normalized data, using the R package DESeq based on a negative binomial model, followed by a Benjamini and Hochberg correction on the p-values with a threshold of 5% to account for multiple comparisons. To assess differences in expression in wheat leaves, we used the set of 8528 triplets of homoeologs. Each homoeolog contribution is considered, independently, as RPKM normalized reads counts.

Under the assumption that global gene expression (GGE) is contributed by the different copies (homoeologs), we calculate GGE as the sum of A_h, B_h and D_h (or $D_{t)}$ homoeolog expressions in the natural cv Courtot or the synthetic TC109 allohexaploid wheats; the sum of A_h and B_h homoeolog expression (HE) in its extracted allotetrapoid "Tetra-Courtot" and equal to D_t expression in the diploid *Ae. tauschii*. Data normalization and statistical analysis steps were performed using R platform (Team, 2008). We compared both global and homoeolog gene expression changes in the different wheat species at different ploidy levels.

Results

Decreasing and re-increasing ploidy level, such as in the allohexaploid wheat model used here, involve regulation of both global and homoeologous gene expression.

Out of the 8 605 homoeologous gene triplets, 8 528 (99.1%) showed at least one expressed homoeolog in at least one sample. Only 77 genes (0.9%) showed no expression for all the three homoeologs in all analyzed samples, and were discarded.

Partitioning of global gene expression between constituent homoeologs in the allohexaploid wheat cv Courtot and its extracted allotetraploid wheat "Tetra-Courtot"

Analyses reveal that most of the three homoeologs are generally expressed in the natural wheat allohexaploid cv Courtot (Figure S3.a). We observed 4617 genes (54.1%) where the A_h homoeolog contributes equally to the B_h one (not significantly differentially expressed), 4926 genes (57.7%) to D_h one and 4787 genes (56.1%) where D_h and B_h contribute equally (Figure 2a). For 3234 genes (37.9% of total) all three homoeologs are equally expressed (Figure 2a). We observed nearly similar numbers of genes (1069 for A_h, 1101 for B_h and 1028 for D_h), where one homoeolog was significantly more expressed than the two others homoeologs, which can be, in their turn, either equally or differentially expressed (Figure 2a). These observations indicate the contribution of the three homoeologs to expression of the majority of the genes in the natural allohexaploid wheat and the nearly absence of significant bias towards dominance of one subgenome over the others.

Similarly, 4510 (52.9%) of A_h and B_h homoeologs were equally expressed in the wheat allotetraploid "Tetra-Courtot". The A_h homoeolog was more expressed than the B_h homoeolog for 2044 (24%) genes; at nearly similar level for the inverse situation (1974, 23%, Figure 2b), also indicating the absence of subgenome dominance.

It is thus important to directly compare homoeolog expression between the different species with different ploidy levels but also in comparison to global gene expression, calculated here as the sum of expression of constituent homoeologs.

Comparison of global gene expression between the natural wheat species

Considering GGE, 5819 genes (68.2%) were similarly expressed when comparing the allohexaploid wheat cv Courtot to its extracted allotetraploid "Tetra-Courtot". Almost half (1299 ~47.9% genes) of the 2709 differentially expressed genes are up-expressed in the

natural allohexaploid wheat cv Courtot whereas, the other half (1410 ~52%) are up-expressed in the extracted allotetraploid "Tetra-Courtot" (Figure 3).

GGE in the diploid specie *Ae. tauschii*, donor of the D genome, is the same here as the estimation at the homoeolog level. Comparison with GGE estimated in the natural allohexaploid wheat cv Courtot shows that 55.1% of the 8528 genes (4697) are equally expressed (Figure 3). The level of differentially expressed genes between the two species is higher than what was observed using microarrays (Chague et al., 2010; Chelaifa et al., 2013 ; Chalabi et al., 2014). For the remaining differentially expressed genes, the proportion of genes that are more expressed in the allohexaploid wheat cv Courtot (2278, 26.7%) was higher than that in *Ae. tauschii* (1552, 18.2%) (Figure 3).

There was also a higher number of genes more expressed in the extracted allotetraploid "Tetra-Courtot" (2545, 29.8%) than in *Ae. tauschii* (1923, 22.5%) whereas majority of the remaining 4060 (47.6%) were equally expressed (Figure 3).

The extracted allotetraploid wheat "Tetra-Courtot" was closer to natural wheat allotetraploid cv Joyau than to any other wheat species with 6848/8528 ~80% equally expressed genes (Table 1). Interestingly, there was two time more expressed genes in Tetra-Courtot (1091) than in Joyau (589) when comparing other extracted and natural wheat allotetraploids (Table 1), as also observed when using microarray (Chague et al., 2010; Chelaifa et al., 2013 ; Chalabi et al., 2014) as well as with a different group (Zhang et al., 2014).

Comparison of homoeolog gene expression between the natural wheat species

It is important to dissect global expression of genes into that of constituent homoeologs and compare across the different wheat species.

Comparison of A_h and B_h homoeolog expression

The expression ratio of A_h or B_h homoeologs in the extracted allotetraploid "Tetra-Courtot" over that in the natural allohexaploid cv Courtot shows a distribution with a peak of 3 to 2 (3/2) (Figure 4.a). This indicates a general trend of lower expression of A_h and B_h homoeologs in the wheat allohexaploid cv Courtot compared to that of its extracted allotetraploid wheat "Tetra-Courtot". To further confirm this, we applied a statistical test and showed that majority of A_h (8090~94.9%) and B_h (7961~93.4%) homoeologs are expressed in the natural wheat allohexaploid cv Courtot at levels equal to 2/3 the level of expression in the extracted allotetraploid "Tetra-Courtot" (Figure 5a). For the majority of these, A_h and B_h

homoeologs belong to the same genes (7673/8528~90%) (Figure 5a). This 2/3 relationship is true for the majority of genes (5819) that show equal GGE as well as for those genes up- or down-expressed in the allohexaploid cv Courtot as compared to its extracted allotetraploid "Tetra-Courtot" (Figure 5a).

Consequently, direct comparison between the two allopolyploids reveals 3311~38.8% A_h and 3283~38.5% B_h homoeologs that are significantly up-expressed in the extracted B_hA_h allotetraploid "Tetra-Courtot" as compared to its natural allohexaploid cv Courtot (Figure 6), 2297 of which belongs to the same genes. Only few A_h (67) and B_h (68) homoeologs (7 from same genes) were less expressed in the extracted B_hA_h "Tetra-Courtot" as compared to the natural allohexaploid cv Courtot (Figure 6).

For those 1410 genes with a GGE higher in the extracted tetraploid "Tetra-Courtot" than the natural allohexaploid cv Courtot, (825/1410) ~58.5% of A_h and B_h homoeoalleles are simultaneously up-expressed in the extracted tetraploid "Tetra-Courtot" (Figure 7a). For remaining genes the up-expression of one homoeolog (A_h or B_h) in the extracted tetraploid wheat goes with an equal expression level of the other homoeolog (Figure 7a).

For the 1299 genes with up-GGE in the natural allohexaploid wheat cv Courtot as compared to its extracted A_hB_h tetraploid "Tetra-Courtot", almost half (598/1299~46%) exhibit an equal expression level for A_h and B_h homoeologs; whereas 186/1299~14.3% exhibit an up-expression of both A_h and B_h homoeologs in the extracted tetraploid "Tetra-Courtot" (Figure 7c). Majority of the remaining ((200+217)/1299~32.1%) exhibit up-expression of one homoeolog in the extracted allotetraploid "Tetra-Courtot" whereas the other one has equal expression level as in the wheat allohexaploid cv Courtot (Figure 7c). Only few genes showed in the extracted wheat allotetraploid "Tetra-Courtot" lower expression of both homoeologs (7), or lower expression of one homoeolog with simultaneously equal expression of the second one (66) (Figure 7c). This suggests that the up GGE in the allohexaploid wheat cv Courtot is most-likely contributed by the D_h genome for the majority of these genes (*see below*).

The general trend of 3/2 up expression of A_h and B_h homoeologs in the extracted allotetraploid "Tetra-Courtot", compared to the natural allohexaploid cv Courtot, suggests a compensation up-expression when the D_h genome was eliminated.

For the 80% of genes with equal GGE, majority of constituent A_h and B_h homoeologs were also equally expressed between the two wheat allotetraploids "Tetra-Courtot" and Joyau (Table 6). Among genes more expressed in Tetra-Courtot than in Joyau, for 394/1091 ones

both A_h and B_h homoeologs are more expressed in "Tetra-Courtot" than their equivalent A_j and B_j homoeologs in cv; Joyau (Table 1) whereas for 225 and 401 genes only A_h and B_h homoeologs are respectively more expressed. Among the genes more expressed in Joy, for 92/589 ones both A and B homoeologs of Joyau are more expressed than A_h and B_h ones in "Tetra-Courtot" whereas only one homoeolog was more expressed for 179 A_j and 257 B_j (Table 1).

Comparison of expression between the D homoeolog

While 4697~55.1% of the genes show equal GGE in the natural wheat allohexaploid and *Ae. tauschii*, distribution of ratios of D_t gene expression in *Ae. taushii* over that of D_h homoeolog in the natural allohexaploid wheat cv Courtot exhibits a peak of density at 3 to one (Figure 4.a). Statistical tests show, as expected, that majority of these (6499~76.2%) are less expressed in the wheat allohexaploid as compared to the diploid wheat whereas another test shows that for (7489)~87.8% D_h homoeologs are expressed in allohexaploid wheat cv Courtot at a level equal to 1/3 that of their corresponding D_t in *Ae. tauschii*.

Reprogramming of gene expression when reestablishing allohexaploidy

Removing the D_h genome from the natural allohexaploid cv Courtot had led to the extracted wheat tetraploid 'Tetra-Courtot'. Obtaining synthetic allohexaploid TC109, by hybridizing the allotetraploid "Tetra-Courtot" (B_hA_h) with the diploid progenitor *Ae. tauschii* (Mestiri et al., 2010), consists in adding again the D_t genome and reincreasing ploidy level. This constitutes a unique system in order to check how the different homoeologs, whose expression had changed when extracting the allotetraploid "Tetra-Courtot," will behave in the synthetic wheat allohexaploid TC109 where the B_hA_h genome is joined to the D_t genome. Various comparisons are done here at both GGE and individual homoeolog expression levels.

Comparison of GGE between natural and synthetic wheat allohexaploids

The majority of the genes (8050, ~94.4%), showed equal global gene expression between natural and synthetic allohexaploid TC109 except 245 and 233 genes that are up- and down-expressed in the synthetic allohexaploid TC109, respectively (Figure 8).

Comparisons also reveal that for the majority of the genes that became differentially expressed in the extracted allotetraploid "Tetra-Courtot" (1280 out of 1410 up-expressed genes, ~90.8 % and 1105 out of 1299, down-expressed ~85.1% genes) readopt when adding

the D_t genome in the synthetic wheat allohexaploid TC109, similar expression levels to those of original natural wheat allohexaploid cv Courtot (Figure 8). For the remaining, there were 149 and 122 genes respectively down and up-expressed in the extracted wheat allotetraploid that didn't readopt in the synthetic allohexaploid TC109 GGE equal to those of natural allohexaploid wheat cv Courtot (Figure 8).

Similarly, the majority of the genes equally- (4589/4697~97.7%), up- (1405/1552~90.5%) and down-expressed (2056/2279~90.2%) when comparing natural allohexaploid cv Courtot to D_t genome donor *Ae. tauschii* are equally expressed when comparing natural and synthetic allohexaploid wheats (Table 2).

Comparison of homoeologous expression between natural and synthetic wheat allohexaploids

As for GGE, the majority of A_h (8374~98.2%), B_h (8296~97.3%), and a lower proportion of D_h homoeologs (7838~91.9%) are equally expressed in the synthetic TC109 and the natural wheat cv Courtot allohexaploids (Tables 3, 4). Cross comparison of global gene expression and that of their constituent homoeologs show several interesting features:

(i) The equal GGE in both the natural cv Courtot and synthetic wheat TC109 allohexaploids was also the case for their three A_h, B_h and $D_{h/t}$ homoeologs for the majority of the genes 7528/8050~93.5% (Table 4). For the remaining genes, we observe significantly higher or lower expression of one homoeolog and rarely of two, without affecting estimation of global gene expression that remains equal (Table 4). The majority of these concerns D homoeologs with 301 and 85 genes, where the D_h homoeolog in the natural wheat allohexaploid cv Courtot was respectively more and less expressed than the D_t homoeolog in synthetic allohexaploid TC109 wheat whereas A_h and B_h homoeologs remain equally expressed (Table 4). This would reflect a higher divergence of D_t and D_h homoeologs as compared to A_h and B_h homoeologs hich are theoretically the same in tetraploid and hexaploid wheats.

(ii) The higher global gene expression of the 233 genes in the natural wheat allohexaploid cv Courtot than in the synthetic one TC109 could be inferred to higher expression of the D_h homoeolog for 132, the A_h homoeolog for 27, the B_h homoeolog for 28, the B_h and D_h homoeologs for 15, the A_h and D_h homoeologs for 15 and all three homoeologs for only 4 genes (Table 4).

(iii) The lower global gene expression of the 245 genes in the natural wheat allohexaploid cv Courtot than in the synthetic one (TC109) could be inferred to lower expression of the D_h homoeolog for 83, the A_h homoeolog for 27 one, the B_h homoeolog for 59, the A_h and B_h homoeologs for 7, the A_h and D_h homoeologs for 14, the B_h and D_h homoeologs for 11, and finally all three homoeologs for 14 (Table 4).

These observations allow us to conclude that the situations where GGE could be equal whereas homoeologs could be differentially expressed between the two wheat allohexaploids, in a compensation manner, are insignificant or rare. Also D homoeologs show the most divergent expression and explain the majority of those genes with differential global gene expression between the two wheat allohexaploids, certainly because of the different origins and thus higher divergence of D_h and D_t homoeologs.

Regulation of homoeolog expression at different ploidy levels

The fact that the majority of homoeologs are equally expressed in synthetic TC109 and natural allohexaploid wheat cv Courtot (Tables 3, 4), suggests an interesting up- and down-expression regulation when decreasing and reincreasing ploidy levels.

The majority of homoeologs increase expression when separated and decrease expression when joined together

We checked that most of the genes for which A_h and B_h homoeologs became up-expressed in the extracted allotetraploid wheat 'Tetra-Courtot' compared to the natural allohexaploid cv Courtot (3209/3283~97.7% for A_h and 3167/3311~95.6% for B_h respectively), decrease homoeolog expression and readopt in the synthetic allohexaploid TC109 similar homoeolog expression to that in the natural wheat allohexaploid cv Courtot (Tables 3.a and 3.b).

Similarly, most of the D_t genes that were up-expressed in the D genome donor diploid species *Ae. tauschii* as compared to their corresponding D_h homoeologs in natural allohexaploid wheat cv Courtot (6499/8528~76.2%) decrease expression and readopt in the synthetic wheat allohexaploid TC109 equal expression patterns to that of natural allohexaploid wheat cv Courtot (6017/6499~92.6%) (Table 3.c).

These observations reveal very interesting features in this unique model of decreasing and reincreasing polyploidy level:

(i) The increasing expression of A_h and B_h homoelog expressions in the extracted allotetraploid "Tetra-Courtot", where the D_h genome was eliminated. The higher

expression of D_t homoeolog in the diploid *Ae.tauschii* compared to its expression level in allohexaploids.

(ii) The decreasing expression of A_h and B_h homoeologs when joined again with the D_t subgenome, through the synthetic allohexaploid TC109, to levels similar to their original expression in the natural allohexaploid wheat cv Courtot. This is concomitantly accompanied by expression in the synthetic allohexaploid TC109 of D_t homoeologs at levels similar to natural wheat allohexaploid cv Courtot, compensating A_h and B_h expression decrease and maintaining the two allohexaploids at similar GGE.

Therefore these homoeologs increase expression when separated, through decreasing ploidy level and decrease expression when joined together, through increasing ploidy level.

Few homoeologs decrease expression when separated and increase when joined together

Interestingly, 27/38 A_h and 27/67 B_h homoeologs (with only 2 belonging to common genes), of the rare cases that became down-expressed in the extracted allotetraploid readopt in the synthetic allohexaploid wheat TC109 similar expression to that of the natural allohexaploid cv Courtot (Tables 3.a and 3.b). This suggests that up-expression of these homoeologs in natural allohexaploid cv Courtot is induced or regulated by the D_h homoeolog. Similarly, we also found 7 D_t homoeologs that were down expressed in the D genome donor diploid species *Ae. tauschii* as compared to the natural allohexaploid wheat cv Courtot and where expression increased in the synthetic allohexaploid wheat TC109 to equal levels to those in the natural one, suggesting that the higher expression is probably regulated by the A_h and/or B_h homoeologs.

Homoeologs adopting new expression profiles in synthetic wheat allohexaploid TC109

We dissected homoeologs that did not readopt an expression level in synthetic allohexaploid TC109 similar to that of natural wheat allohexaploid cv Courtot.

We found 11 A_h and 15 B_h homoeologs that became up-expressed in the synthetic allohexaploid TC109 compared to both natural wheat allohexaploid cv Courtot and its extracted wheat tetraploid "Tetra-Courtot" (Tables 3a and 3b). This suggests that D_t homoeologs in the synthetic allohexaploid TC109 (which are different from D_h homoeoalleles in the natural specie) may have induced this up-expression of A_h and B_h homoeologs in the synthetic wheat allohexaploid TC109. Similarly, we also found 4 D_t

homoeologs that became up-expressed in the synthetic allohexaploid TC109 compared to both natural wheat allohexaploid cv Courtot and the diploid D genome donor *Ae. tauschii* (Table 3c).

For the reverse situation, there are 28 A_h and 33 B_h homoeologs that are down-expressed in the synthetic wheat allohexaploid TC109 compared to natural allohexaploid wheat cv Courtot and its extracted tetraploid "Tetra-Courtot", with only 2 from same genes (Tables 3a and 3b, Table 5). This up- and down-expression of A_h and B_h homoeologs suggest a regulation mechanism most likely induced by the D_t homoeologs. We also found 115 D_t homoeologs that are down-expressed in the synthetic wheat allohexaploid TC109 compared to both natural wheat allohexaploid cv Courtot and the diploid D genome donor *Ae. tauschii* (Table 3c).

Homoeologs that do not accommodate expression when re-increasing ploidy levels

There were 69 A_h and 138 B_h homoeologs that became up-expressed in the extracted tetraploid "Tetra-Courtot" and continue to be up-expressed in the synthetic allohexaploid TC109 and didn't readopt a similar expression to natural wheat allohexaploid cv Courtot (Table 3.a and 3.b), 15 belonging to the same genes (Table 5). Similarly, 207 homoeologs up-expressed in D diploid specie *Ae. tauschii* continue to be up-expressed in the synthetic allohexaploid TC109 and didn't readopt a similar expression to natural wheat allohexaploid cv Courtot (Table 3c). Inversely, there were 8 A_h and 38 B_h homoeologs that became down-expressed in the tetraploid "Tetra-Courtot" and continue to be down-expressed in the synthetic allohexaploid TC109 as compared to natural wheat allohexaploid cv Courtot (Table 3.a and 3.b). This is also the case for 89 D_t homoeologs down-expressed in the D diploid species *Ae. tauschii* and synthetic allohexaploid TC109 compared to the natural wheat allohexaploid cv Courtot.

For 275 genes, the D homoeologs are up-expressed in diploid species *Ae. tauschii* compared to natural allohexaploid cv Courtot, whereas they are up-expressed in the natural allohexaploid cv Courtot compared to synthetic one TC109 (Table 3c).

Discussion

The evolution of gene expression is a driving force for phenotype diversification in all organisms. Recent studies have shown that allopolyploidizations induce rapid and evolutionary changes in gene expression (Wendel, 2000; Adams et al., 2003; Hegarty et al., 2006 ; Wang et al., 2006a ; Chaudhary et al., 2009 ; Pumphrey et al., 2009 ; Rapp et al., 2009 ; Akhunova et al., 2010 ; Chague et al., 2010 ; Buggs et al., 2011b ; Grover et al., 2012 ; Li et al., 2014 ; Rambani et al., 2014; Zhang et al., 2014). The reprogramming of gene expression, may represent an adaptive mechanism for a stable polyploid species (Paterson et al., 2011 ; Rambani et al., 2014 ; Xu et al., 2014). It has been suggested that these changes could be due to interactions between parental regulatory networks, stoichiometric disruptions due to the incongruence between WGD and gene dosage, and *de novo* genetic and epigenetic alterations (Osborn et al., 2003 ; Riddle and Birchler, 2003 ; Comai, 2005 ; Adams and Wendel, 2005b ; Chen, 2007 ; Doyle et al., 2008 ; Soltis and Soltis, 2009 ; Jackson and Chen, 2010 ; Birchler, 2012 ; Madlung and Wendel, 2013).

Most studies of gene expression changes undertaken up to day were indirect i.e. mostly comparing global gene expression between polyploids and their parents or average of parents without dissecting at the single duplicated copies (homoeologs). Our understanding of gene expression regulation at the homoeolog scale as well as at the global expression scale was possible here by the availability of the 8605 triplet genes of hexaploid wheat, with sequences of all three homoeologs (IWGSC, 2014), combined with the original wheat polyploid model with decreasing and re-increasing ploidy levels (Chague et al., 2010 ; Mestiri et al., 2010; Chelaifa et al., 2013). The wheat allohexaploidy is decreased here by eliminating the D_h genome and disengaging the allotetraploid B_hA_h genome "Tetra-Courtot" that cohabitated with the D_h genome for several thousand years and generations. Ploidy level is re-increased and allohexaploidy is reestablished by adding the D_t genome through the production of synthetic wheat allohexaploids TC109.

We showed here that partitioning of gene expression between constituent homoeologs is largely established in natural allohexaploid wheat, its extracted wheat allotetraploid as well as the newly-synthesized allohexaploid wheat; with patterns of both 'genome dominance' and 'genome equivalence'. Interestingly, the majority of homoeologs contribute equally to gene expression, whereas for the remaining, nearly similar proportions of biased expression

towards homoeologs from one specific subgenome are observed in all three species (Figure 2). The absence of significant bias towards any subgenome, in the recent wheat allohexaploid, but also in the older wheat allotetraploid, is in concordance with partitioning expression observed in the recent *Brassica napus* allopolyploid (Chalhoub et al., 2014) and hexaploid wheat (IWGSC, 2014). This contrasts with many old and recent polyploids (Buggs et al., 2011b; Schnable et al., 2011; Ilut et al., 2012; Tang et al., 2012; Yoo et al., 2013) but concurs with other old polyploids (Garsmeur et al., 2014).

Interestingly, 31.8% of the 8528 genes became differentially expressed, at nearly similar proportions of up- and down-expressed genes, when decreasing ploidy level in the extracted allotetraploid "Tetra-Courtot", giving a primary indication that decreasing ploidy levels does not lead to a global decrease in gene expression. The majority of genes (8050, ~94.4%), reestablish in the synthetic allohexaploid wheat TC109 similar global gene expression levels to those of natural allohexaploid wheat cv Courtot.

When dissected at the constituent homoeolog level, we observe that the majority of the homoeologs are up-expressed in the extracted allotetraploid "Tetra-Courtot" and the diploid *Ae. tauschii* whereas they decrease their expression and readopt when joined together in the newly-synthesized allohexaploids wheat equal expression to the natural one (Tables 3, 4).

Thus a main finding of the present study is that the majority of homoeologs generally increase expression when separated and decrease expression when joined again together. This is demonstrated here for 97.7%, 95.7% and 92.6% of respectively the 3283 A_h, 3311 B_h and 6499 D_t homoeologs that are up-expressed in tetraploid and diploid wheats (Tables 3, 4) whereas they decrease their expression when joined together in synthetic allohexaploid wheat TC109 to similar levels of that of the natural wheat allohexaploid cv Courtot (Figure 4, Table 3). The increased and decreased expression in opposite direction to ploidy levels is also demonstrated by the fact that expression of A_h and B_h homoeologs in natural and resynthesized allohexaploid wheat are generally 2/3 of their expression in the extracted allotetraploid 'Tetra-Courtot', whereas those of the D_h and D_t homoeologs in allohexaploids are 1/3 of that of the D_t genes in the diploid *Ae. tauschii* (Figure 4).

For the remaining genes that are not equally expressed at the global level between natural and synthetic wheat allohexaploids, our data suggest that they can largely explained by divergent expression between D_t and D_h homoeologs for 57.1% for those up-expressed in

the natural allohexaploid and 33.9% for those down-expressed (Table 4), certainly because of the different origins of the D_h and D_t homoeologs.

Patterns of expression regulation that do not follow this rule of increased and decreased expression inversely to ploidy levels and adopt novel expression patterns were incipient. These include few homoeologs that decrease expression when separated and increase expression when joined together; those that adopt divergent expression only when reincreasing ploidy level (Tables 3, 4). We generally found a higher proportion of D_t homoeologs that adopt new expression profiles than A_h and B_h homoeologs, illustrating again the divergence between D_h and D_t genomes.

Dissecting gene expression as done in the present study also shows that situations where GGE could be equal whereas homoeologs could have a "certain degree of freedom" to vary their expression in a compensation manner between the different polyploids are rare.

Interestingly, a higher proportions of up-expressed genes is confirmed here in the extracted allotetraploid "Tetra-Courtot" as compared to the natural *T. turgidum spp durum* cv Joyau, in agreement with our previous study using microarrays (Chalabi et al., 2014) as well as another study using also microarrays with different extracted and natural wheat (Zhang et al., 2014) with a third different extracted wheat allotetraploid. Our study showed that the upexpression in Tetra-Courtot, is attributable to both homoeologs for 394/1091~36.1% of the genes, to A_h homoeolog for 215~19.7% and to B_h one for 401~36.7% genes. It has been suggested that this could be explained by a possibility that B_hA_h component of allohexaploid wheat would have accumulated extensive and functionally distinct heritable changes (Zhang et al., 2014). However, the fact that both microarray and RNA-Seq technologies as used here are based on reference sequences available for the $B_hA_hD_h$ genome of hexaploid wheat (IWGSC, 2014) may explain that higher expression values for its extracted allotetraploid and Dh homoeologs would be rather more detected than the divergent natural wheat allotetraploid cv Joyau.

Nevertheless, these different comparisons suggest that trans-regulation (repression) of A_hB_h subgenome by the D_h subgenome is established in natural allohexaploid wheat rather than a cis-acting regulation. Disengaging the D_h subgenome from the B_hA_h subgenomes, after 10000 years of cohabitation in the natural allohexaploid wheat cv Courtot, leads to up-expression of an important proportion of A_h and B_h homoeologs that are immediately repressed upon joining them together with the D_t subgenome in the synthetic allohexaploid wheat.

Observations about decrease and increase expression of wheat homoeologs inversely to that of ploidy levels is in agreement with the gene dosage balance hypothesis (Veitia et al., 2008; Birchler and Veitia, 2012). The dosage balance hypothesis predicts that central network genes, which have many key interactions with other metabolic components, should be preferentially retained because elimination of such a gene would disturb the stoichiometry of many interactions. Dosage balance considers the stoichiometry of synthetized molecules to maintain a function, such as, in our case of changes in ploidy level: an increase in gene copy number induces a decrease of homoeolog expression level to maintain a function, inversely when decreasing gene copy number, the increase in homoeolog expression level ensures gene function. It would be very important to dissect those A_h, B_h and D_h homoeologs that do not follow in hexaploid wheat the dosage partitioning and genome equivalence rule.

While it has been suggested that transcriptome changes in extracted tetraploid wheat (Zhang et al., 2014) could be due to gene loss (Brenchley et al., 2012) from B and A subgenomes (Pont et al., 2013), genetic alterations, such as gene conversion, copy number variation, and transposition (Saintenac, 2011), alternative splicing (Akhunova et al., 2010); (Akhunov et al., 2013) and heritable epigenetic alterations (Shaked et al., 2001 ; Zhao et al., 2011), it is important to note that by focusing here on the triplets we targeted genes that were not or less affected by such changes. It will be thus interesting to compare expression of those genes present as doublets or single copies in these wheat genomes.

Results obtained in this work contribute to our understanding of gene expression regulation at different ploidy levels by dissecting the global gene expression at the level of constituent homoeologs. The future functional analysis of the different gene expression categories would reveal important gene functional categories that are regulated in response to polyploidy.

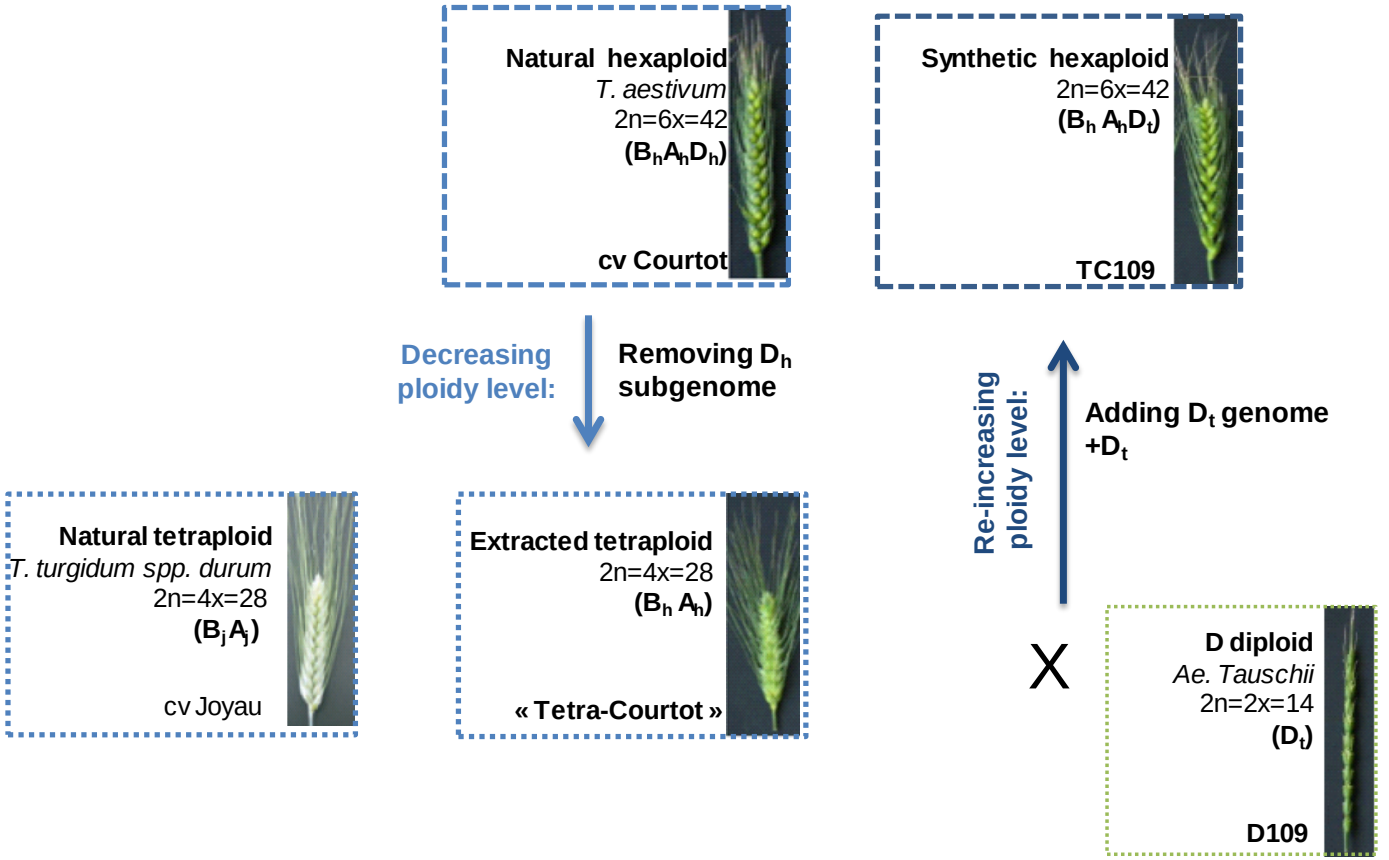


Figure 1: Schematic representation of origin of the extracted tetraploid and the synthetic wheat allohexaploid (details see Mestiri et al, 2010). The tetraploid component «Tetra-Courtot » (2n=4x= 28, B_hA_h) is extracted from the natural allohexaploid wheat cv Courtot (2n=6x= 42, $B_hA_hD_h$) (Mestiri et al, 2010). The newly-synthetized wheat allohexaploid were obtained through hybridization between the extracted tetraploid «Tetra-Courtot» and *Aegilops tauschii* accession AtD109 (2n=2x= 14, D_t) followed by spontaneous chromosome doubling. Here, photos represente spike of the natural allohexaploid cv Courtot, the natural tetraploid *T.turgidum* cv Joyau, the extracted allotetraploid « Tetra-Courtot », the diploid specie *Ae.tauschii* and the synthetic allohexaploid TC109. A_h, B_h, D_h, A_j, B_j, D_t correspond to homoeologs, where j subscript denotes Joyau, h for hexaploid and t for *tauschii*.

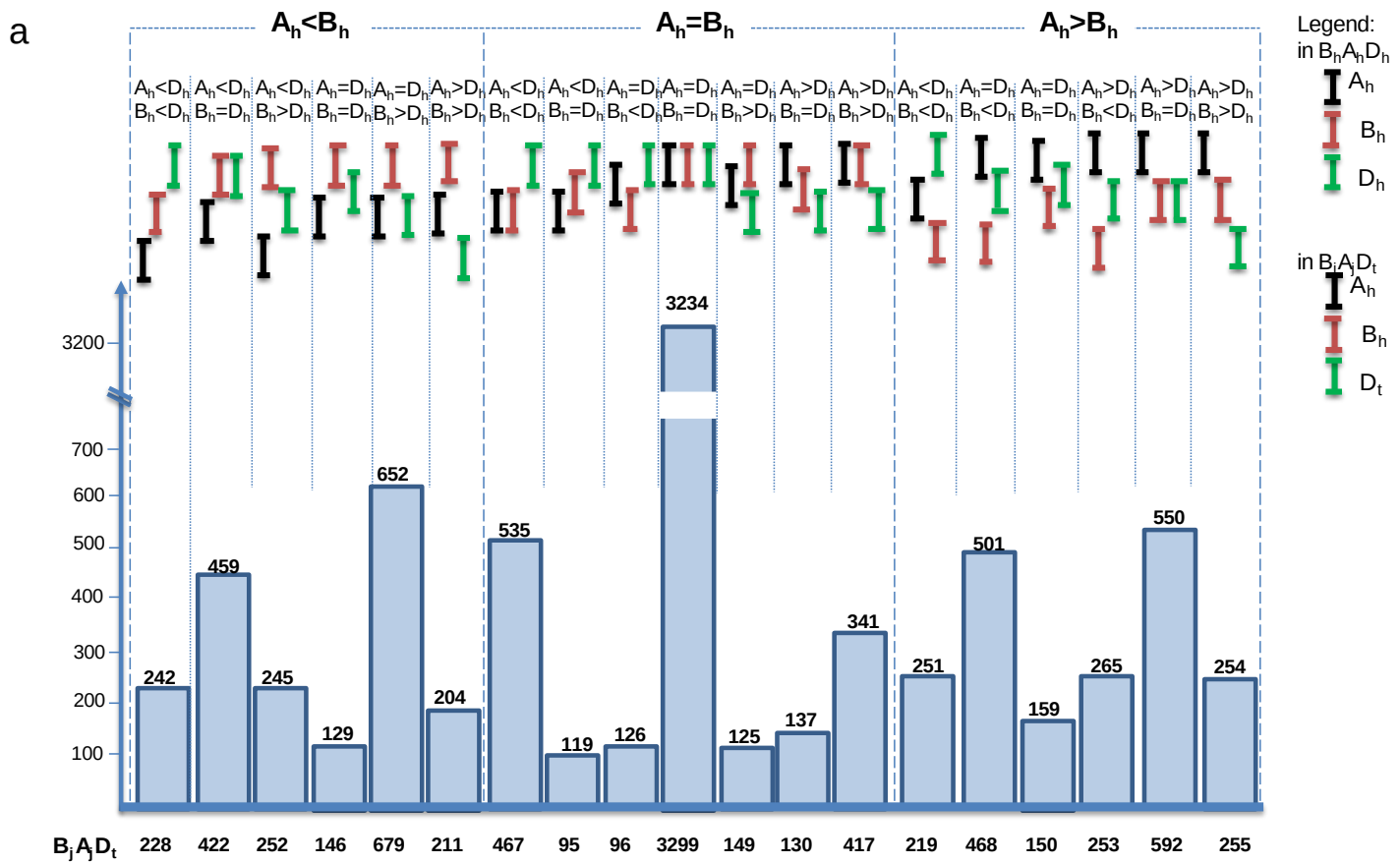


Figure 2: **Expression patterns revealed when comparing homoeologs expression levels. a.** Expression patterns revealed when comparing A_h, B_h and D_h homoeologs expression levels in the natural wheat allopolyploid cv Courtot $B_hA_hD_h$. Black bars depict the expression level of A_h homoeologs, red bars those of B_h homoeologs, and green bars those of the D_h homoeologs. Expression is differential between two homoeologs when their expressions (bars) do not overlap whereas it is equal when they do overlap. The y axis represents gene count. The x axis represents expression patterns.The number of genes for each expression pattern in each of the expression categories is indicated to help comparisons. Values observed in the newly synthetized allohexaploid $B_jA_jD_t$ are presented below the histogram. In the newly synthesized allohexaploid, homoeologs are A_h, B_h and D_t. Differential expression has been calculated by RPKM normalization and DESeq differential analysis (see Materials and Methods).

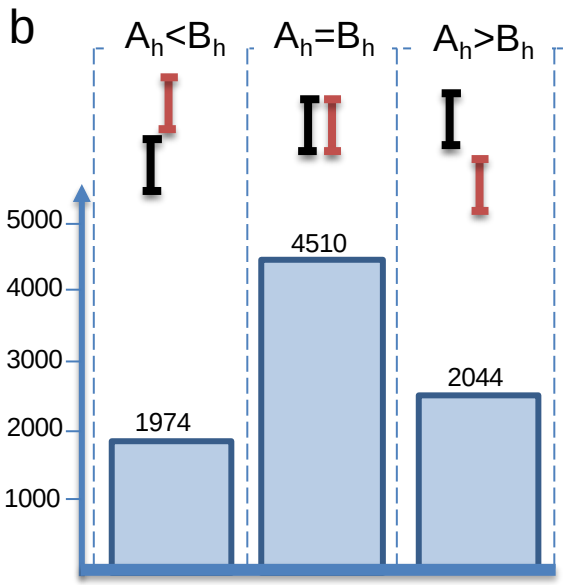


Figure 2: **b.** Expression patterns revealed when comparing A_h and B_h homoeologs in the extracted allotetraploid «Tetra-Courtot» B_hA_h. Black bars depict the expression level observed of A_h homoeologs, and red bars depict those B_h.

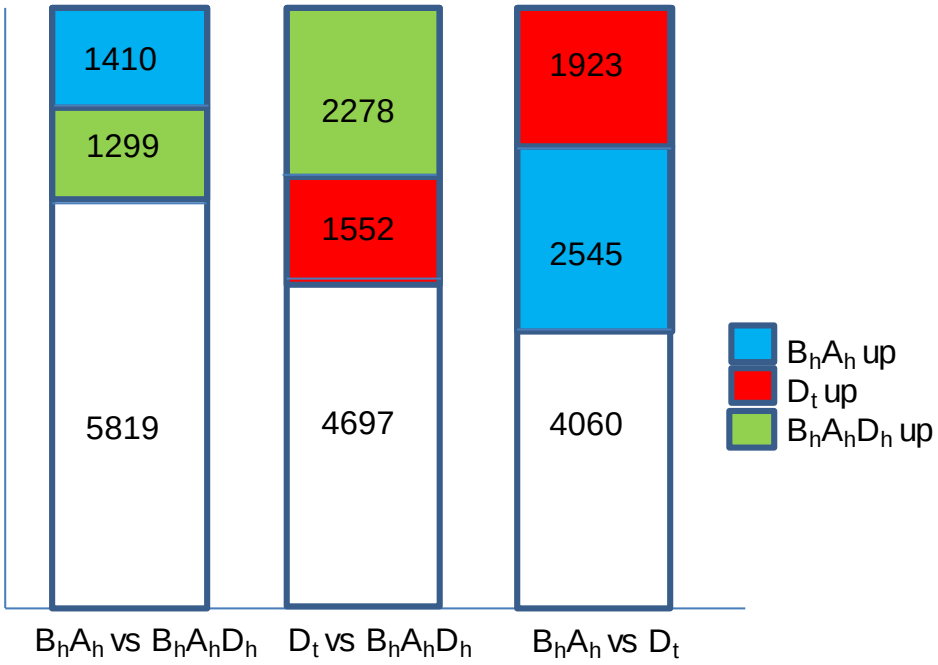


Figure 3: Schematic representation of global gene expression comparison between different wheat species : B_hA_h denote the extracted tetraploid « Tetra-Courtot », D_t denote the D_t diploid D genome donnor *Ae.tauschii*, and $B_hA_hD_h$ the natural allohexaploid cv Courtot. Blue color corresponds to an up-expression of B_hA_h, red an up-expression of D_t, green to an up-expression in $B_hA_hD_h$, and white color to an equal global gene expression.

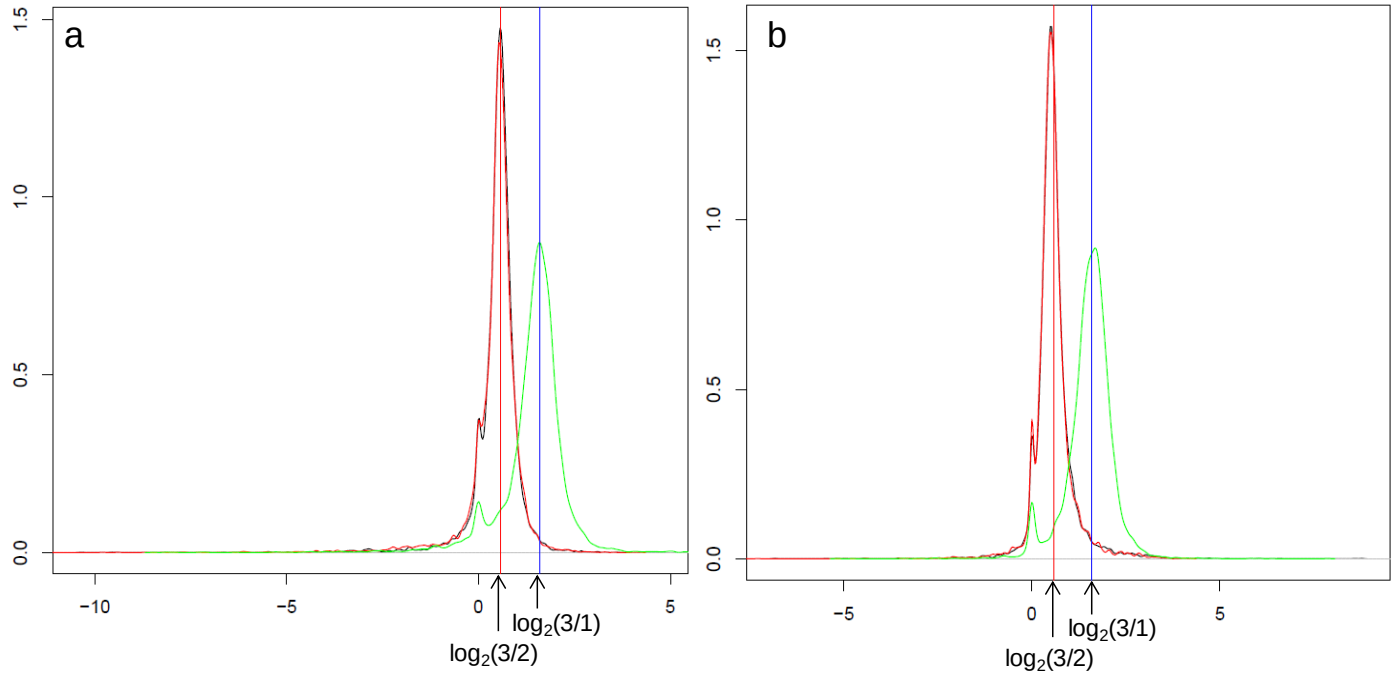


Figure 4: Density distribution of ratios of homoeologs expression level.

(a) Ratios of A_h, B_h homoeolog expression level in the tetraploid «Tetra-Courtot» B_hA_h (black and red, respectively) and that of D_t expression in the diploid *Ae.tauschii* over D_h homoeolog in the natural allohexaploid wheat cv Courtot $B_hA_hD_h$ (green).

(b) Ratios of A_h, B_h homoeolog expression level in the tetraploid «Tetra-Courtot» B_hA_h (black and red, respectively) and that of D_t expression in the diploid *Ae.tauschii* over the same D_t in the newly-synthesized allohexaploid $B_hA_hD_t$.

Red and blue lines vertical lines bring out 3/2 and 3/1 ratio values.

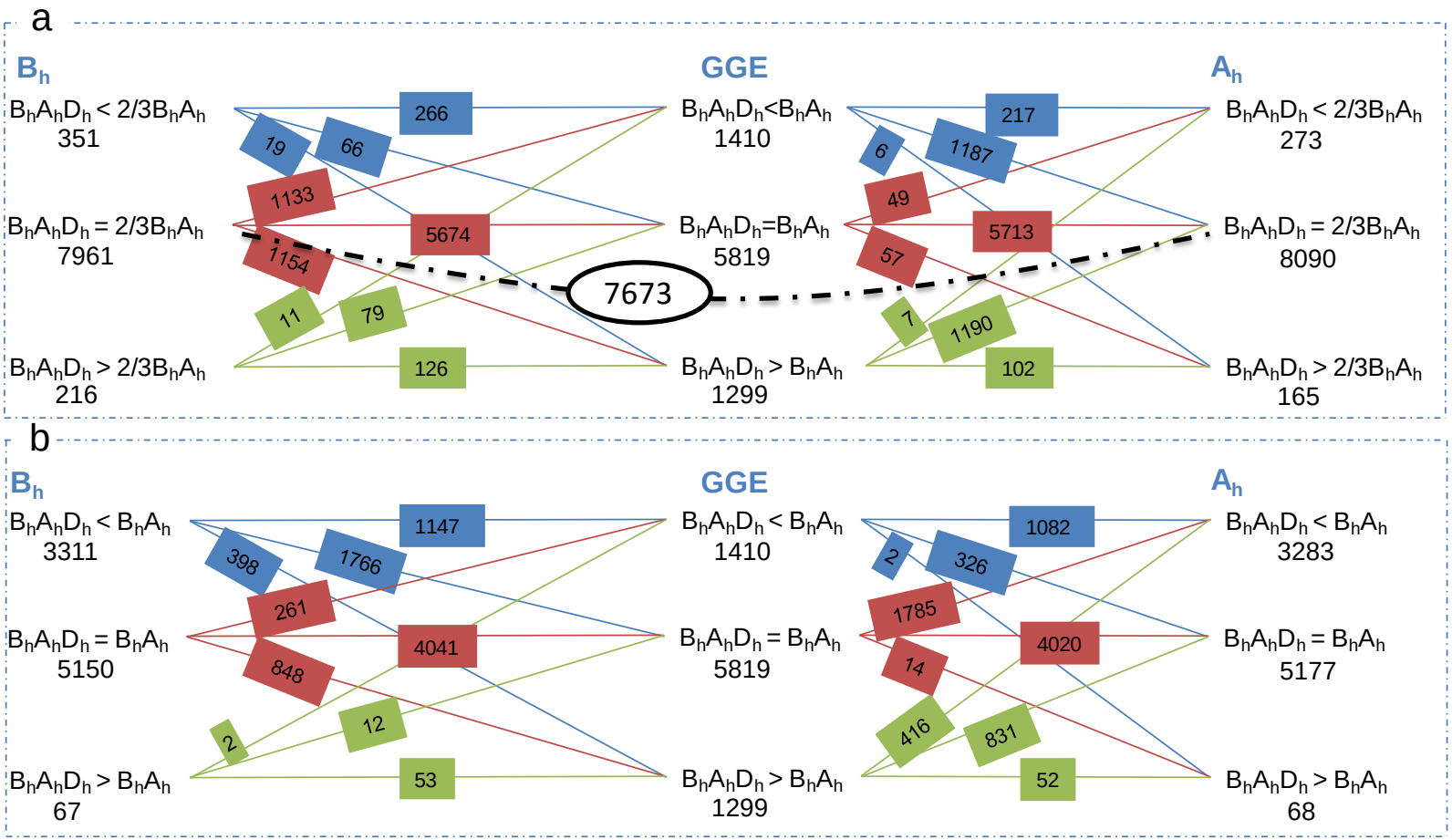


Figure 5: Cross comparison between categories of expression revealed by comparing global gene expression (GGE), A_h and B_h homoeologs expressions between natural wheat allohexaploid cv Courtot ($B_hA_hD_h$) and its extracted allotetraploid « Tetra-Courtot » (B_hA_h). a) Comparison of A_h and B_h homoeologs expression in cv Courtot to 2/3 values of their expression in « Tetra-Courtot » **. b)** Comparison of A_h and B_h homoeologs expression in cv Courtot to direct values of their expression in « Tetra-Courtot » Numbers of genes of different categories are presented below each categories. Numbers of genes shared by different expression categories of global gene expression and homoeologous expression are indicated on the cross-lines. Numbers of genes shared by 2 categories of homoeologs A_h and B_h are surrounded. B_hA_h, allotetraploid "Tetra-Courtot"; $B_hA_hD_h$, natural allohexaploid cv Courtot.

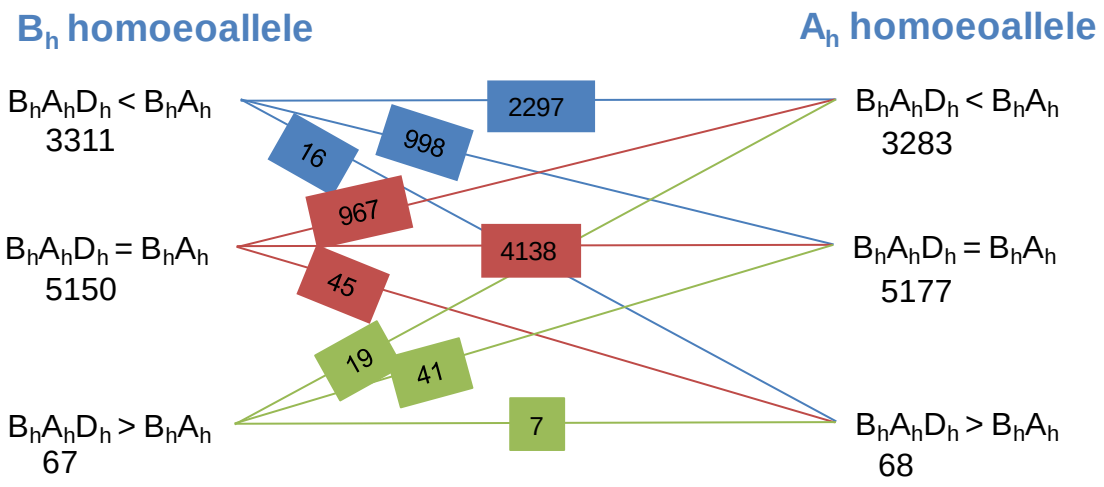


Figure 6: Comparison of sets of A_h and B_h homoeolog expression between the natural allohexaploid cv Courtot ($B_hA_hD_h$) and its extracted « Tetra-Courtot » (B_hA_h). Numbers of genes of different categories are presented below each category. Numbers of genes shared by the different expression categories of homoeologs expression are indicated on the cross-lines.

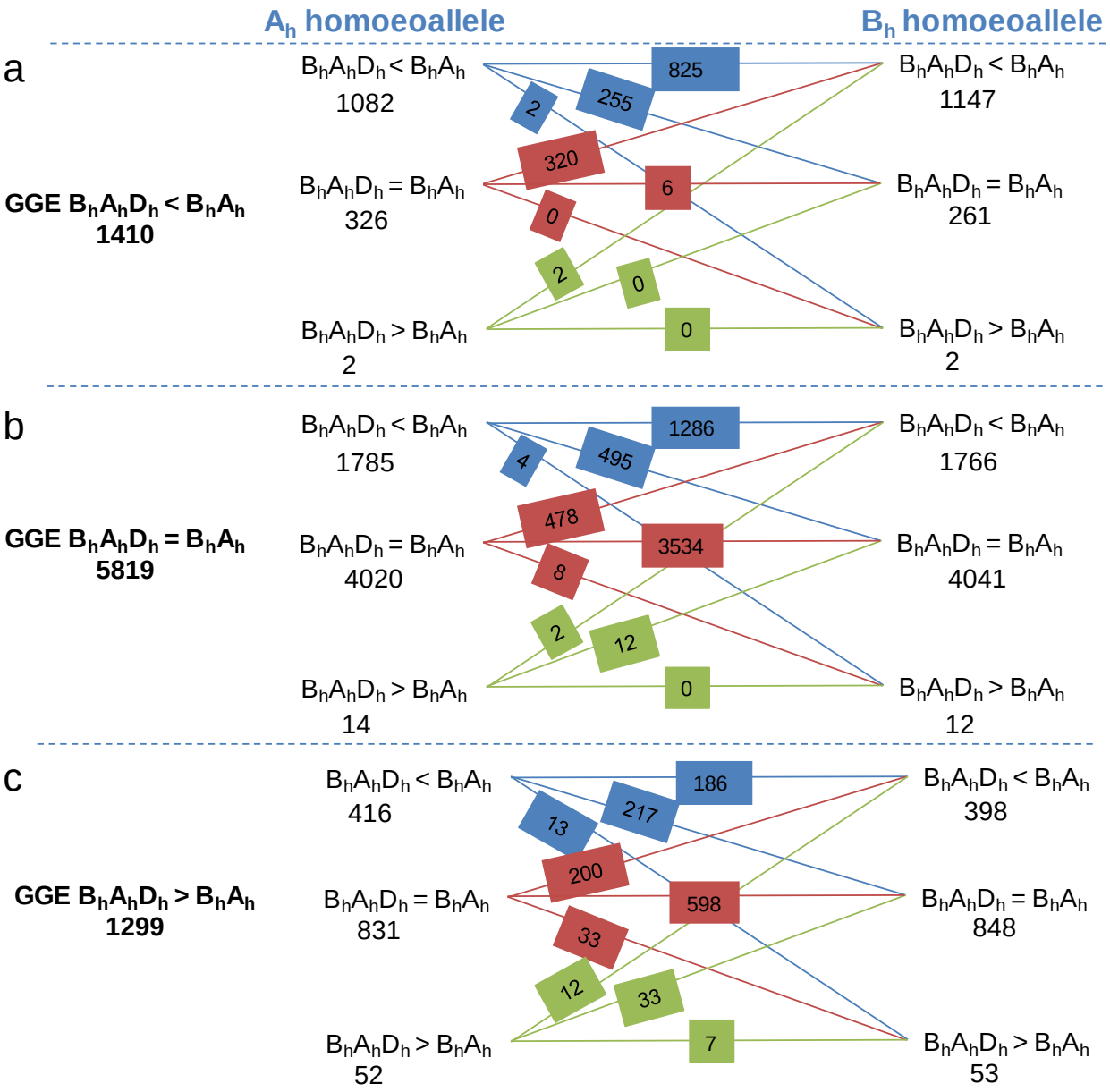


Figure 7: Comparison of A_h and B_h homoeologs expression between the natural allohexaploid cv Courtot $B_hA_hD_h$ and its extracted « Tetra-Courtot » B_hA_h crossed with global gene expression (GGE) comparison. Numbers of genes shared by the different expression categories of homoeolog expression are indicated on the cross-lines. Numbers of genes of different categories are presented below each category. B_hA_h, allotetraploid "Tetra-Courtot"; $B_hA_hD_h$, natural allohexaploid cv Courtot.

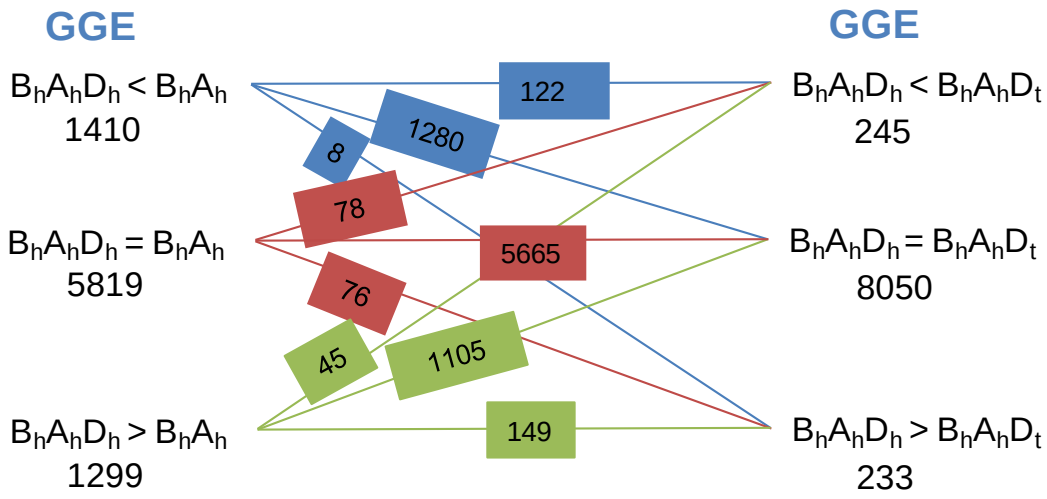


Figure 8: Comparison of sets of differentially and equally expressed transcripts, at the global gene expression (GGE) level, revealed between natural allohexaploid cv Courtot ($B_hA_hD_h$) and its extracted tetraploid « Tetra-Courtot » (B_hA_h) from one side, and the natural allohexaploid cv Courtot ($B_hA_hD_h$) and the newly synthesized allohexaploid $B_hA_hD_t$ from the other side. Numbers of genes shared by the different expression categories of global gene expression and homoeologous expression are indicated on the cross-lines. Numbers of genes of different categories are presented below each categories. B_hA_h, allotetraploid "Tetra-Courtot"; $B_hA_hD_h$, natural allohexaploid cv Courtot ; $B_hA_hD_t$, the newly synthesized allohexaploid TC109.

Table 1: Cross comparisons of A_h and B_h homoeologs expression and global gene expression between the natural allotetraploid cv Joyau (B_jA_j), the allotetraploid « Tetra-Courtot » (B_hA_h).

GGE B_jA_j vs B_hA_h	B_jA_j vs B_hA_h <			B_jA_j vs B_hA_h =			B_jA_j vs B_hA_h >			Total
	<	=	>	<	=	>	<	=	>	
<	394	215	10	401	66	0	5	0	0	1091
=	0	84	11	229	6284	136	9	95	0	6848
>	0	0	8	0	49	257	4	179	92	589
Total	394	215	29	630	6399	393	18	274	92	8528

Table 2: Cross comparisons of global gene expression between the natural allohexaploid cv Courtot ($B_hA_hD_h$), the diploid *Ae.tauschii* (D_t), and the synthetic allohexaploid TC109 ($B_hA_hD_t$).

	$B_hA_hD_h<D_t$	$B_hA_hD_h=D_t$	$B_hA_hD_h>D_t$	Total
$B_hA_hD_h < B_hA_hD_t$	78	58	109	245
$B_hA_hD_h = B_hA_hD_t$	2056	4589	1405	8050
$B_hA_hD_h > B_hA_hD_t$	145	50	38	233
Total	2279	4697	1552	8528

Table 3: Cross comparisons of A_h, B_h, D_h, D_t homoeologs expression level between the natural allohexaploid cv Courtot ($B_hA_hD_h$) and the synthetic allohexaploid TC109 ($B_hA_hD_t$), and Courtot ($B_hA_hD_h$) and its extracted tetraploid tetra-Courtot (B_hA_h) for A_h, B_h homoeologs as well as the D_h homoeolog in Courtot to that of the D_t one in the diploid *Ae.tauschii*.

	« Tetra-Courtot » (B_hA_h) compared to natural allohexaploid cv Courtot ($B_hA_hD_h$)								the diploid *Ae.tauschii* (D_t) compared to natural allohexaploid cv Courtot ($B_hA_hD_h$)			
	a A_h homoeolog B_hA_h vs $B_hA_hD_h$				b B_h homoeolog B_hA_h vs $B_hA_hD_h$				c D homeolog D_t vs $B_hA_hD_h$			
	<	=	>	Total	<	=	>	Total	<	=	>	Total
$B_hA_hD_h < B_hA_hD_t$	3	11	69	83	2	15	138	155	0	4	207	211
$B_hA_hD_h = B_hA_hD_t$	27	5138	3209	8374	27	5102	3167	8296	7	1814	6017	7838
$B_hA_hD_h > B_hA_hD_t$	8	28	5	41	38	33	6	77	89	115	275	479
Total	38	5177	3283	8528	67	5150	3311	8528	96	1933	6499	8528

Table 4: Cross comparisons of global gene expression and A_h, B_h, D_h and D_t homeologs expression between the natural allohexaploid cv Courtot ($B_hA_hD_h$) and the synthetic allohexaploid TC109 ($B_hA_hD_t$).

The natural allohexaploid cv Courtot ($B_hA_hD_h$) compared to the newly synthetized allohexaploid TC109 ($B_hA_hD_t$)

HE	$B_hA_hD_h < B_hA_hD_t$									$B_hA_hD_h = B_hA_hD_t$									$B_hA_hD_h > B_hA_hD_t$								
	$B_hA_hD_h<B_hA_hD_t$			$B_hA_hD_h=B_hA_hD_t$			$B_hA_hD_h>B_hA_hD_t$			$B_hA_hD_h<B_hA_hD_t$			$B_hA_hD_h=B_hA_hD_t$			$B_hA_hD_h>B_hA_hD_t$			$B_hA_hD_h<B_hA_hD_t$			$B_hA_hD_h=B_hA_hD_t$			$B_hA_hD_h>B_hA_hD_t$		
	D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t			D_h vs D_t		
GGE	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>
$B_hA_hD_h<B_hA_hD_t$	14	7	0	14	27	0	0	1	0	11	59	0	83	28	0	1	0	0	0	0	0	0	0	0	0	0	0
$B_hA_hD_h=B_hA_hD_t$	0	0	0	0	18	1	0	0	0	0	55	9	85	7528	301	1	26	0	0	0	0	1	24	0	0	0	0
$B_hA_hD_h>B_hA_hD_t$	0	0	0	0	0	1	0	0	0	0	0	0	1	10	132	0	28	25	0	0	0	0	27	15	0	0	4
Total	14	7	0	14	45	2	0	1	0	11	114	9	169	7566	433	2	54	16	0	0	0	1	51	15	0	0	4

Table 5: Cross comparisons of A_h and B_h homoeologs expression between the natural allohexaploid cv Courtot ($B_hA_hD_h$), the allotetraploid « Tetra-Courtot » (B_hA_h) and the synthetic allohexaploid TC109 ($B_hA_hD_t$).

A_h HE		$B_hA_h < B_hA_hD_h$			$B_hA_h = B_hA_hD_h$			$B_hA_h > B_hA_hD_h$			
B_h HE		$B_hA_hD_h < B_hA_hD_t$	$B_hA_hD_h = B_hA_hD_t$	$B_hA_hD_h > B_hA_hD_t$	$B_hA_hD_h < B_hA_hD_t$	$B_hA_hD_h = B_hA_hD_t$	$B_hA_hD_h > B_hA_hD_t$	$B_hA_hD_h < B_hA_hD_t$	$B_hA_hD_h = B_hA_hD_t$	$B_hA_hD_h > B_hA_hD_t$	
$B_hA_h < B_hA_hD_h$	$B_hA_hD_h < B_hA_hD_t$	1	0	0	0	1	0	0	0	0	2
	$B_hA_hD_h = B_hA_hD_t$	1	2	2	0	20	0	0	2	0	27
	$B_hA_hD_h > B_hA_hD_t$	0	0	1	0	20	0	0	17	0	38
$B_hA_h = B_hA_hD_h$	$B_hA_hD_h < B_hA_hD_t$	0	0	0	4	8	0	1	2	0	15
	$B_hA_hD_h = B_hA_hD_t$	1	21	23	5	4078	18	27	928	1	5102
	$B_hA_hD_h > B_hA_hD_t$	0	0	0	0	23	**2**	1	7	0	33
$B_hA_h > B_hA_hD_h$	$B_hA_hD_h < B_hA_hD_t$	0	0	0	0	56	0	**15**	67	0	138
	$B_hA_hD_h = B_hA_hD_t$	0	4	12	2	929	8	25	2184	3	3167
	$B_hA_hD_h > B_hA_hD_t$	0	0	0	0	3	0	0	2	1	6
		3	27	38	11	5138	28	69	3209	5	8528

Table 6: Cross comparisons of A_h ,B_h homoeologs expression and global gene expression between the natural allotetraploid cv Joyau (B_jA_j), the allotetraploid « Tetra-Courtot » (B_hA_h).

		A_j vs A_h									
		<			=			>			
		Bj vs B_h			Bj vs B_h			Bj vs B_h			
		<	=	>	<	=	>	<	=	>	Total
GGE B_jA_j vs B_hA_h	<	394	215	10	401	66	0	5	0	0	1091
	=	0	84	11	229	6284	136	9	95	0	6848
	>	0	0	8	0	49	257	4	179	92	589
	Total	394	215	29	630	6399	393	18	274	92	8528

Table S1: Plant materials used in this study and statistics of RNA-Seq read counts obtained for each replicate and reads mapped by BWA (Li et al, 2010).

Notation	Species	Ploidy level	Nb total of examined reads	Nb total of mapped reads	% of single aligned reads
D_t	Ae.tauschii 109 -11	diploid	39765951	4005265	10.07
D_t	Ae.tauschii 109 -14	diploid	74803165	6113986	8.17
D_t	Ae.tauschii 109 -16	diploid	83298821	6072181	7.29
B_jA_j	Joyau 11-2	tetraploid	24158972	3082484	12.76
B_jA_j	Joyau 12-2	tetraploid	21956365	2976091	13.55
B_jA_j	Joyau 15-1	tetraploid	19851102	2579384	12.99
B_hA_h	Tetra-Courtot BC6 -9	tetraploid	74342414	8028397	10.79
B_hA_h	Tetra-Courtot BC6 -13	tetraploid	68074206	7224715	10.61
B_hA_h	Tetra-Courtot BC6 -14	tetraploid	73274610	7392371	10.08
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	71398502	8386094	11.74
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	73576289	8401191	11.41
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	82785477	8593258	10.38
$B_hA_hD_h$	Courtot	hexaploid	84639474	9160678	10.82
$B_hA_hD_h$	Courtot	hexaploid	81965434	8436605	10.29
$B_hA_hD_h$	Courtot	hexaploid	82617976	8708512	10.54

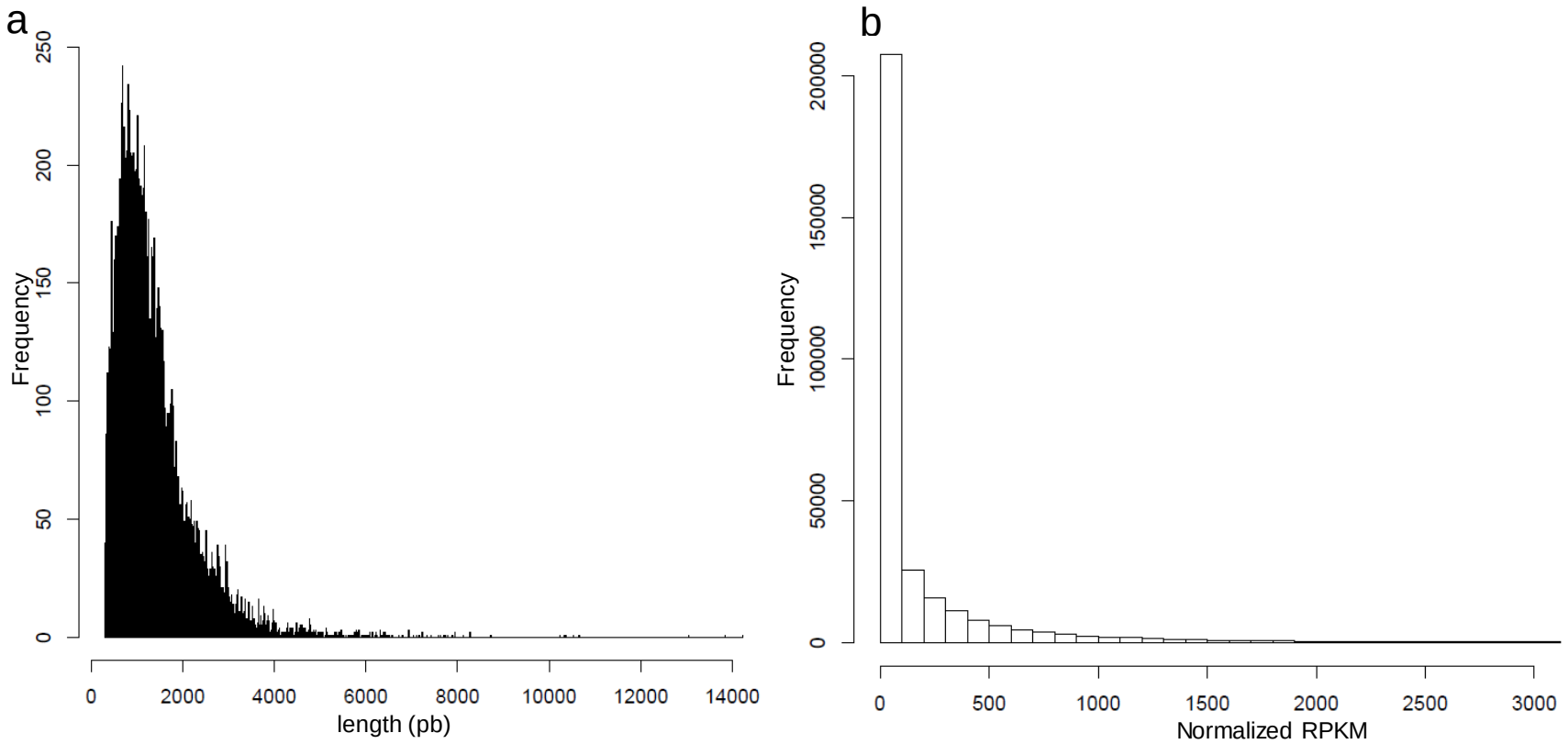


Figure S1. Distribution of the 8528 triplets of wheat genes (IWGSC et al, 2014) used in the present study by the length of available sequence **(a)**, and by their normalised RPKM counts **(b)**.

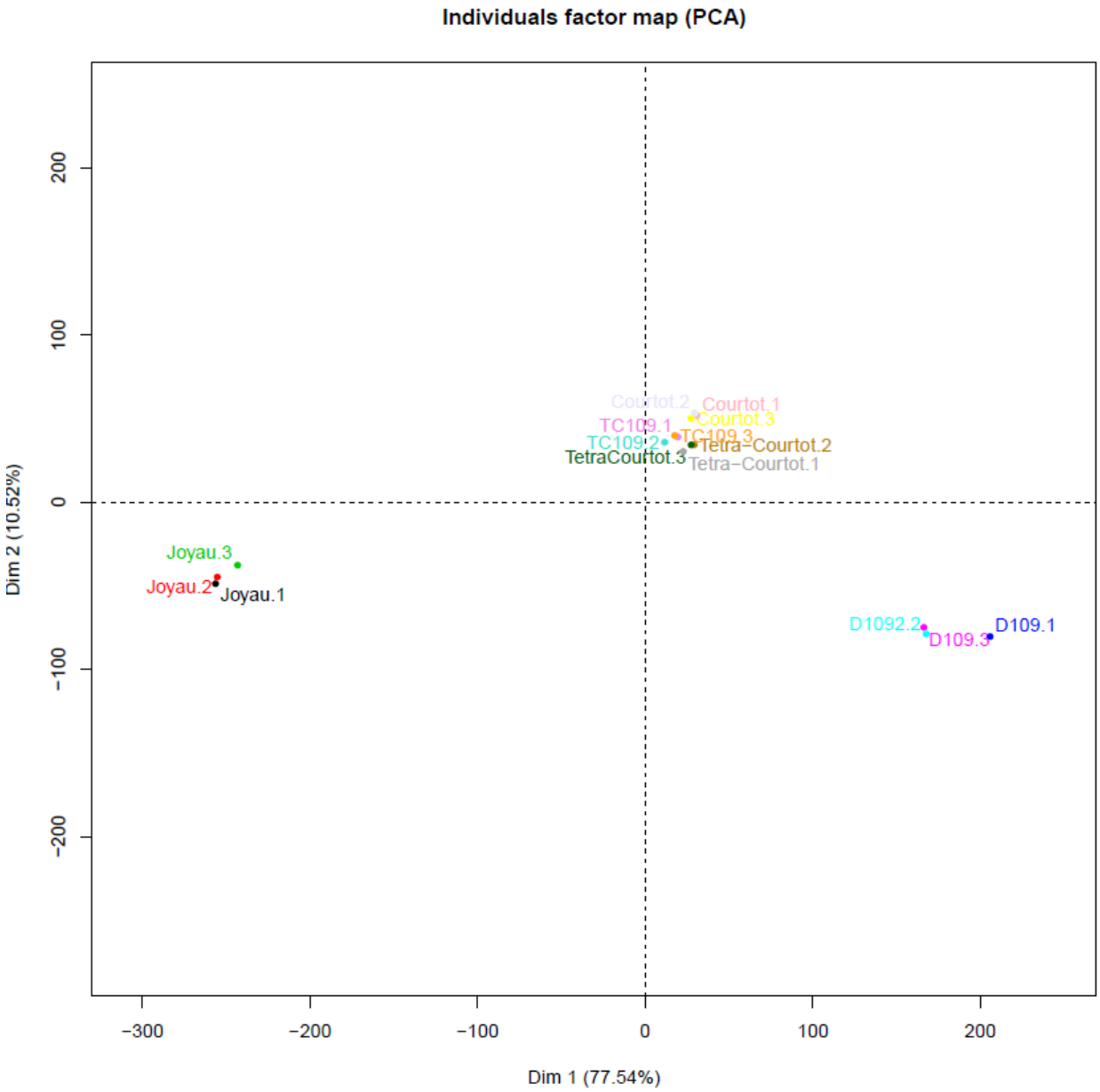


Figure S2: Principal component analysis (PCA) applied to summarize the expression data and identify spurious technical effects. The projection of the 15 samples on the first two PC-axes in the sample space shows a satisfactory reproducibility between biological replicates, with those two axes explaining almost 88.06% of the variance.

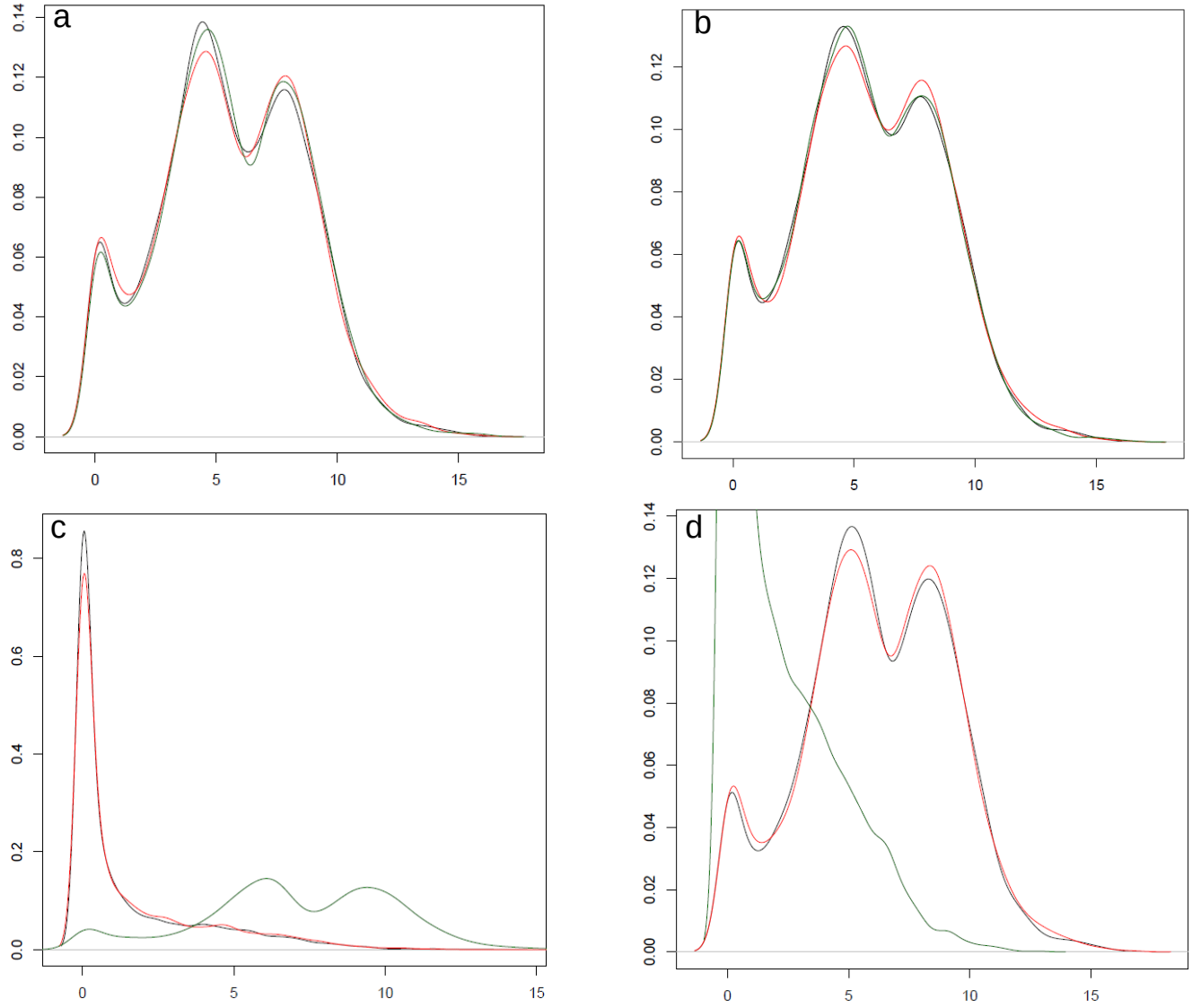


Figure S3: Density of distribution of RNA sequence reads mapped on A_h, B_h and D_h homoeologs (normalized read counts) in natural allohexaploid wheat cv Courtot $B_hA_hD_h$ (**a**), the newly synthetized allohexaploid TC109 $B_hA_hD_t$(**b**), the D_t diploid genome donor *Ae.tauschi*i (**c**), the extracted tetraploid « Tetra-Courtot » B_hA_h (**d**). Homoeologs are represented in black color for A_h, red for B_h, and green for D_h or D_t.

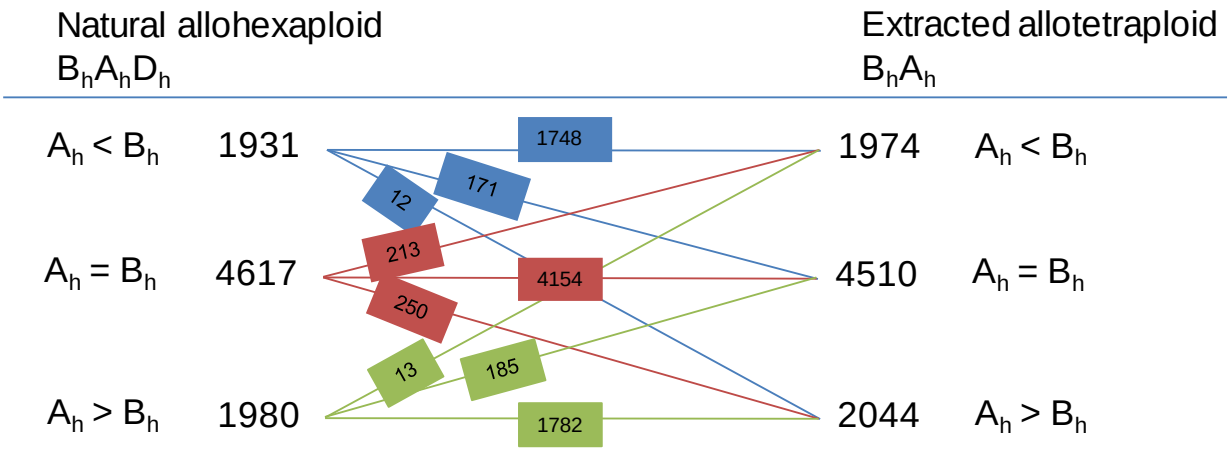


Figure S4: Cross comparison of A_h and B_h homoeologs expression level in the natural allohexaploid $B_hA_hD_h$ and in the extracted allotetraploid B_hA_h. Numbers of genes shared by the different expression categories of the natural allohexaploid $B_hA_hD_h$ and the extracted B_hA_h are indicated on the cross-lines.

References

Adams, K., Cronn, R., Percifield, R., and Wendel, J. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ specific reciprocal silencing. Proc Natl Acad Sci USA 100, 4649-4654.

Adams, K.L., and Wendel, J. (2004). Exploring the genomic mysteries of polyploidy in cotton. Biological Journal of the Linnean Society 82, 573-581.

Adams, K.L., and Wendel, J.F. (2005b). Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8, 135-141.

Akhunov, E.D., Sehgal, S., Liang, H., Wang, S., Akhunova, A.R., Kaur, G., Li, W., Forrest, K.L., See, D., Simkova, H., Ma, Y., Hayden, M.J., Luo, M., Faris, J.D., Dolezel, J., and Gill, B.S. (2013). Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol 161, 252-265.

Akhunova, A.R., Matniyazov, R.T., Liang, H., and Akhunov, E.D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics 11, 505.

Bell, G.D., Kane, N.C., Rieseberg, L.H., and Adams, K.L. (2013). RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. Genome Biol Evol 5, 1309-1323.

Birchler, J.A. (2012). Genetic Consequences of Polyploidy in Plants. In Polyploidy and Genome Evolution., P.S. Soltis and D.E. Soltis, eds (Berlin Heidelberg: Springer-Verlag), pp. 415.

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

Blake, N.K., Lehfeldt, B.R., Lavin, M., and Talbert, L.E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome 42, 351-360.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G., D'Amore, R., Allen, A., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., Kay, S., Waite, D., Trick, M., Bancroft, I., Gu, Y., Huo, N., Luo, M., Sehgal, S., Gill, B., Kianian, S., Anderson, O., Kersey, P., Dvorak , J., McCombie, W., Hall, A., Mayer, K., Edwards, K., Bevan, M., and Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. 491, 705-710.

Buggs, R.J., Zhang, L., Miles, N., Tate, J.A., Gao, L., Wei, W., Schnable, P.S., Barbazuk, W.B., Soltis, P.S., and Soltis, D.E. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol 21, 551-556.

Chague, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., and Chalhoub, B. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol 187, 1181-1194.

Chalabi, S., Chelaifa, H., Arnaud D, LeFloch E, Mestiri, I., DinhThi, V., LeClainche, I., Belcram, H., Devauchelle, C., Rizzon, C., Deffains, D., Huteau, V., Coriton, O., Chiquet, J., Jahier, J., and Chalhoub, C. (2014). Unraveling gene expression changes when decreasing and re-increasing alloploidy in wheat. in prepartion.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans B, Corréa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle

G, Mestiri I, Schnel N, Le Paslier MC, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VH, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CH, Wang X, Canaguier A, Chauveau A, Bérard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Wang X, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, and Wincker, P. (2014). Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. Science 345, 950-953

Chalupska, D., Lee, H.Y., Faris, J.D., Evrard, A., Chalhoub, B., Haselkorn, R., and Gornicki, P. (2008). Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci U S A 105, 9691-9696.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M., and Wendel, J.F. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). Genetics 182, 503-517.

Chelaifa, H., Chague, V., Chalabi, S., Mestiri, I., Arnaud, D., Deffains, D., Lu, Y., Belcram, H., Huteau, V., Chiquet, J., Coriton, O., Just, J., Jahier, J., and Chalhoub, B. (2013). Prevalence of gene expression additivity in genetically stable wheat allohexaploids. New Phytol 197, 730-736.

Chen, Z.J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58, 377-406.

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

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

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

Dvorak, J., Terlizzi, P., Zhang, H.B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome 36, 21-31.

Flagel, L.E., and Wendel, J.F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol 186, 184-193.

Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., and Osborn, T.C. (2007). Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19, 3403-3417.

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

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homeolog expression bias and expression level dominance in allopolyploids. The New Phytologist. 196, 966-971.

Hegarty, M., Barker, G., Wilson, I., Abbott, R., Edwards, K., and Hiscock, S. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Curr Biol. 16, 1652-1659.

Hegarty, M.J., Jones, J.M., Wilson, I.D., Barker, G.L., Coghill, J.A., Sanchez-Baracaldo, P., Liu, G., Buggs, R.J., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2005). Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. Mol Ecol 14, 2493-2510.

Higgins, J., Magusin, A., Trick, M., Fraser, F., and Bancroft, I. (2012). Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species Brassica napus. BMC Genomics. 13, 247.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc Natl Acad Sci U S A 99, 8133-8138.

Ilut, D.C., Coate, J.E., Luciano, A.K., Owens, T.G., May, G.D., Farmer, A., and Doyle, J.J. (2012). A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am J Bot 99, 383-396.

IWGSC. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.

Jackson, S., and Chen, Z.J. (2010). Genomic and expression plasticity of polyploidy. Curr Opin Plant Biol 13, 153-159.

Kihara, H. (1944). Discovery of the DD-analyser, one of the ancestors of vulgare wheats. Ag. Hort. (Tokyo) 19, 889-890.

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

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

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.

Madlung, A., and Wendel, J.F. (2013). Genetic and epigenetic aspects of polyploid evolution in plants. Cytogenet Genome Res 140, 270-285.

McFadden, E.S., and Sears, E.R. (1946). The origin of Triticum speltoides and its free-threshing hexaploid relatives. J. Hered. 37, 81-89.

Mestiri, I., Chague, V., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Chalhoub, B., and Jahier, J. (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. New Phytol 186, 86-101.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods 5, 621-628.

Nesbitt, M., and Samuel, D. (1996). From the staple crop to extinction? The archaeology and history of hulled wheats. . . In Hulled Wheats. Proceedings of the First International Workshop on Hulled Wheats. (International Plant Genetic Resources Institute, Rome, Italy.).

Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., and Martienssen, R.A. (2003). Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19, 141-147.

Page, J.T., Gingle, A.R., and Udall, J.A. (2013). PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 (Bethesda) 3, 517-525.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.J., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.H., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z.,

Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., ur Rahman, M., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.K., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A., Zhang, T., Dennis, E.S., Mayer, K.F., Peterson, D.G., Rokhsar, D.S., Wang, X., and Schmutz, J. (2011). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492, 423-427.

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Dolezel, J., Fahima, T., Budak, H., Keller, B., Salvi, S., Maccaferri, M., Steinbach, D., Feuillet, C., Quesneville, H., and Salse, J. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J 76, 1030-1044.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3, 827-837.

Pumphrey, M., Bai, J., Laudencia-Chingcuanco, D., Anderson, O., and Gill, B.S. (2009). Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics 181, 1147-1157.

Rambani, A., Page, J.T., and Udall, J.A. (2014). Polyploidy and the petal transcriptome of Gossypium. BMC Plant Biol 14, 3.

Rapp, R.A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. BMC Biol 7, 18.

Riddle, N.C., and Birchler, J.A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. Trends Genet 19, 597-600.

Riley, R., Unrau, J., and al., e. (1958). Evidence on the origin of the B genome of wheat. J. Hered. 49, 91-98.

Roulin, A., Auer, P.L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R.W., and Jackson, S.A. (2013). The fate of duplicated genes in a polyploid plant genome. Plant J.

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

Schnable, J., Springer, N., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 108, 4069-4074.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and 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.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu Rev Plant Biol 60, 561-588.

Soltis, P.S., and Soltis, D.E. (2012). Polyploidy and Genome Evolution. (Berlin Heidelberg: Springer-Verlag).

Takumi, S., Nasuda, S., and al., e. (1993). Wheat phylogeny determined by RFLP analysis of nuclear DNA. I. Einkorn wheat. Jpn. J. Genet. 68, 73-79.

Talbert, L.E., Blake, N.K., Storlie, E.W., and Lavin, M. (1995). Variability in wheat based on low-copy DNA sequence comparisons. Genome 38, 951-957.

Tang, H., Woodhouse, M., Cheng, F., Schnable, J., Pedersen, B., Conant, G., Wang, X., Freeling, M., and Pires, J. (2012). Altered patterns of fractionation and exon deletions

in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190, 1563-1574.
Team, R.D.C. (2008). R: A language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing).
Udall, J.A., Swanson, J.M., Nettleton, D., Percifield, R.J., and Wendel, J.F. (2006). A novel approach for characterizing expression levels of genes duplicated by polyploidy. Genetics 173, 1823-1827.
Veitia, R.A., Bottani, S., and Birchler, J.A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet 24, 390-397.
Wang, J., Tian, L., Lee, H.S., and Chen, Z.J. (2006a). Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in Arabidopsis allopolyploids. Genetics 173, 965-974.
Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., and Chen, Z.J. (2006b). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics 172, 507-517.
Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol 42, 225-249.
Xu, C., Bai, Y., Lin, X., Zhao, N., Hu, L., Gong, Z., Wendel, J.F., and Liu, B. (2014). Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. Mol Biol Evol 31, 1066-1076.
Yoo, M.J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) 110, 171-180.
Zhang, H., Zhu, B., Qi, B., Gou, X., Dong, Y., Xu, C., Zhang, B., Huang, W., Liu, C., Wang, X., Yang, C., Zhou, H., Kashkush, K., Feldman, M., Wendel, J.F., and Liu, B. (2014). Evolution of the BBAA Component of Bread Wheat during Its History at the Allohexaploid Level. Plant Cell.
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., and Chalhoub, B. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci U S A 108, 18737-18742.
Zhao, N., Zhu, B., Li, M., Wang, L., Xu, L., Zhang, H., Zheng, S., Qi, B., Han, F., and Liu, B. (2011). Extensive and heritable epigenetic remodeling and genetic stability accompany allohexaploidization of wheat. Genetics 188, 499-510.

4.3. Discussion complémentaire

Les données de séquences des gènes de référence

Dans notre étude, les triplets des trois homéologues des gènes sur lesquels sont alignées les lectures restent partiellement séquencés (IWGSC, 2014) et présentent un recouvrement à 90%, ce qui éviterait des fausses cartographies unique des lectures. Toutefois, j'améliorerai ces séquences références en tronquant les extrémités qui ne sont pas chevauchantes entre les trois homéologues, ce qui augmenterait la spécificité d'alignement. Je garderai également un recouvrement minimal de 300pb de longueur de séquence (IWGSC, 2014 ; Pfeifer et al., 2014).

Les méthodes d'alignement des lectures des séquences mRNA

Dans mes travaux de thèse, j'ai utilisé SOAP2 (Annexe 2) et BWA, ces deux logiciels d'alignement de lectures sont réputés pour être les plus rapides. La rapidité de calcul, la précision d'alignement et les besoins en mémoire sont les éléments clés de comparaison des outils d'alignement de lectures (Lindner and Friedel, 2012; Shang et al., 2014).

J'ai comparé mes travaux à ceux réalisés très récemment sur le génome du blé, dans (Choulet et al., 2014 ; IWGSC, 2014 ; Pfeifer et al., 2014) comparant 5 tissus: racine, feuille, grain, tige et épi (IWGSC, 2014), à 3 étapes de développement dans (Choulet et al., 2014) (soit 30 échantillons en duplicats), et un seul tissu (grain, en 4 réplicats) dans (Pfeifer et al., 2014). Les données ont été séquencées, en single-end (101 pb) et en paired-end (2 x 100 pb), avec Illumina HiSeq2000 dans (IWGSC, 2014) et (Choulet et al., 2014 ; Pfeifer et al., 2014) respectivement. Les lectures des séquences mRNA ont été alignées dans ces travaux avec l'outil Tophat2 (Trapnell et al., 2009 ; Langmead and Salzberg, 2012), qui est un pipeline d'alignement de lectures (utilisant Bowtie2) avec la caractéristique d'aligner les lectures aux jonctions exon-exon. C'est un des outils les plus populaires pour repérer les épissages alternatifs, et est très couramment utilisé dans l'analyse des données d'expression. Le point commun à ces 3 aligneurs est qu'ils utilisent la BWT.

Dans (IWGSC, 2014 ; Pfeifer et al., 2014), les lectures ont été alignées sur l'ensemble du génome (IWGSC, 2014)

- dans (IWGSC, 2014), 3 fois, autorisant 0,1 ou 2 mismatch soit entre 34% (pour 0 mismatch) et 71% (pour 2 mismatch) de lectures alignées ;
- dans (Pfeifer et al., 2014), 2 mismatch par alignement de lectures, soit 64-75% de lectures alignées.

Dans (Choulet et al., 2014), les lectures ont été alignées sur le chromosome 3B avec 0 mismatch et 0 splice-mismatch, correspondant à la sélection de lectures s'alignant parfaitement sur le chromosome 3B afin de repérer uniquement les lectures spécifiques des homéologues. Pour chaque échantillon, en moyenne 50 ±11 million de lectures en paired-end ont été alignées sur le chromosome 3B (sans précision sur le pourcentage de lectures alignées).

La différence observée entre les pourcentages de lectures alignées avec la méthode utilisée dans mes travaux et celles dans (IWGSC, 2014) et (Pfeifer et al., 2014) qui utilisent le génome entier pour l'alignement des lectures, pourrait-être liée :

1) A la provenance des données mRNA-seq de variétés et tissus différents: dans mon étude, j'utilise les données feuilles issus le blé allohexaploïde cv Courtot, tandis que les travaux de (IWGSC, 2014) et (Pfeifer et al., 2014) utilisent le blé allohexaploïde cv Chinese Spring. Toutefois, la différence observée entre ces deux variétés de blé, au niveau de l'expression globale, dans le tissu feuille reste faible 14% (1196/8528 gènes). Mais, il est à noter que le séquençage du génome entier du blé a été réalisé sur cette variété Chinese Spring ;
2) à l'outil d'alignement, Tophat alignerait plus de lectures de part une taille minimum des lectures non imposée, ou la caractéristique principale de Tophat est la détection des jonctions exon-exon ;
3) ou encore à la différence des séquences des gènes référence utilisés pour aligner les lectures RNA-Seq. Dans mes travaux j'utilise uniquement les triplets d'homéologues (A, B et D). De ce fait j'ignore les singletons et doublets d'homéologues, qui sont en effet important pour compléter notre analyse d'expression des homéologues. Les travaux de (IWGSC, 2014) et (Pfeifer et al., 2014) alignent les lectures sur l'ensemble du génome du blé.

L'analyse des données

La normalisation RPKM est critiquée (Dillies et al., 2012 ; Rapaport et al., 2013 ; Chalhoub et al., 2014), bien qu'elle soit encore utilisée (Rambani et al., 2014, Chalhoub,et al., 2014). Une faiblesse, qui est due aux normalisations par le total des lectures alignées, est que la représentation de chacun des gènes dépend des niveaux d'expression de tous les autres gènes. Dans les cas où une petite fraction de gènes représente une grande proportion des lectures séquencées (les gènes fortement exprimés), les faibles changements d'expression dans ces gènes fortement exprimés faussent la normalisation des comptages des gènes faiblement exprimés, ce qui peut induire des erreurs dans l'analyse différentielle. Toutefois, dans le cas précis de mon étude où ~10% de lectures sont alignées, cette normalisation reste appropriée et pour révéler les différences d'expression entre les différents niveaux de ploïdie.
Aussi, la division par la longueur des gènes induit un biais sur les variances: pour un niveau d'expression donné, les gènes les plus courts ont une variance plus grande que les gènes les plus longs.
Après normalisation RPKM, les comptages de lectures ne présentent pas une distribution discrète type Poisson ou Binomiale Négative tel qu'on s'attendrait pour ce type de données à valeurs entières, discrètes, positives. Les données deviennent à valeurs non-entières, ce qui fait que les analyses différentielles basées sur les comptages ne seraient plus adaptées (Soneson and Delorenzi, 2013).

Dans mes travaux, après normalisation de la table des comptages, je sélectionne les triplets de gènes qui sont exprimés pour au moins un homéologue dans au moins un échantillon étudié (Chalhoub et al., 2014 ; Choulet et al., 2014; IWGSC, 2014).

L'analyse différentielle a été réalisée avec DESEQ (Anders and Huber, 2010) (qui fait l'hypothèse d'une loi Binomiale Négative) plutôt qu'une normalisation en log2(RPKM+1) suivie d'un t-test. Cette dernière approche a été testée dans mes travaux, mais les résultats creusent les différences (plus de gènes différentiellement exprimés). En plus, pour des faibles comptages on observe beaucoup plus de différences qu'avec DESEQ qui les considère non significativement différentiellement exprimés. Le t-test ou l'analyse de la variance sont des tests permettant de repérer les différences pour les petits comptages ce qui explique les écarts creusés.

Est-ce qu'il fallait préférer le test t à l'analyse différentielle DESEQ? En comparant les niveaux d'expression dans les génomes du blé naturel allohexaploïde cv Courtot au blé

allotétraploïde « Tetra-Courtot », le t-test montre seulement (3265/8528)~38.3% de gènes à niveau d'expression similaire, et (4965/8528)~58.2% de gènes sur-exprimés dans le naturel allohexaploïde cv Courtot, les 3.5%(298/8528) de gènes restant sont surexprimés dans l'allotétraploïde extrait « Tetra-Courtot ». Ainsi il ne me semblait pas qu'il fallait choisir un test (le test t) qui donne seulement 38% de gènes exprimés de façon équivalente entre Courtot et Tetra-Courtot, ce qui n'est pas comparable avec les analyses basées sur les microarrays (cf Chapitre 2). Grâce à la particularité de DESEq d'être plus conservative (Soneson and Delorenzi, 2013), ce qui signifie que DESEq contrôle très bien le taux de faux positifs, nos résultats montrent des différences entre l'allohexaploïde naturel et l'allotétraploïde extrait « Tetra-Courtot », qui s'approchent plus de ceux obtenus utilisant les approches microarrays.

Troisième Partie

Chapitre 5

Discussion Générale & Perspectives

Au cours de mes travaux de thèse que j'ai entrepris pendant quatre années, j'ai pu profiter des avancées technologiques et scientifiques importantes pour essayer de comprendre la reprogrammation de l'expression des gènes dans les blés polyploïdes.

Ainsi, j'ai bénéficié de l'élaboration et de la caractérisation d'un modèle blé original, permettant de réduire puis d'augmenter le niveau de ploïdie (Mestiri et al., 2010). Toujours dans le but de caractériser la reprogrammation de l'expression des gènes dans ce modèle, j'ai eu l'opportunité d'utiliser les outils microarray disponibles (Chague et al., 2010; Chelaifa et al., 2013).

Les outils d'analyse de l'expression des gènes basés sur les microarrays ne permettant pas de séparer et d'individualiser de façon précise l'expression des différentes copies des gènes dupliquées par allopolyploïdie (ou homéologues), j'ai apporté ma contribution en particulier à ce niveau en développant une composante parent-spécifique de la puce Affymetrix.

Les séquences d'un nombre important d'homéologues ont été, récemment, rendues disponibles à travers le séquençage 'brouillon' du génome du blé allohexaploïde (IWGSC, 2014) et m'ont permises de disséquer la reprogrammation de l'expression des gènes en celles des homéologues qui les composent, utilisant les outils de séquençage RNA-Seq.

En me positionnant par rapport aux travaux entrepris sur d'autres modèles polyploïdes depuis une dizaine d'années (Veitia, 2004; Adams and Wendel, 2005b; Comai, 2005; Madlung et al., 2005; Wendel and Doyle, 2005; Tate et al., 2006; Wang et al., 2006b; Gaeta et al., 2007; Doyle et al., 2008; Leitch and Leitch, 2008; Soltis et al., 2008 ; Akhunova et al., 2010; Chague et al., 2010; Chelaifa et al., 2010b ;Mestiri et al., 2010; Bardil et al., 2011; Brenchley et al., 2012; Combes et al., 2012; Arnaud et al., 2013; Combes et al., 2013; Madlung and Wendel, 2013; Chalhoub et al., 2014; Choulet et al., 2014 ; IWGSC, 2014; Marcussen et al., 2014; Pfeifer et al., 2014; Zhang et al., 2014; Rambani et al., 2014), mes travaux de thèse m'ont permis d'enrichir les différents résultats obtenus et d'approfondir les recherches sur de nombreux aspects de la reprogrammation de l'expression des gènes chez les allopolyploïdes.

5.1. Hypothèse d'additivité et de non-additivité:

Comment révéler les gènes dont l'expression a été reprogrammée dans les polyploïdes ? comparaisons aux parents et à une moyenne parentale.

L'expression des gènes mesurée dans un polyploïde peut être identique ou différente de celle d'un ou des deux parents. L'hypothèse d'additivité a été adaptée à partir de celle utilisée dans la caractérisation des hybrides intraspécifiques par rapport à leur parents ou une moyenne des parents, appelé MPV (Chague et al., 2010). Elle a été appliquée pour la première fois en 2006 en caractérisant la reprogrammation des gènes dans des polyploïdes d'*Arabidopsis* (Wang et al., 2006b) puis largement généralisée et discutée depuis (Arnaud et al., 2013 ; Gianinetti, 2013). Ainsi, selon cette hypothèse, si l'expression dans les polyploïdes est identique à l'un des parents ou différente de celles des deux parents, alors l'expression des gènes est considérée comme 'reprogrammée' (ou non-additive) dans le polyploïde. Si l'expression n'est pas égale à la MPV, on considère que la polyploïdie a eu un effet. Elle est additive dans le cas inverse et on considère que la polyploïdie n'a pas eu d'effet sur la reprogrammation de l'expression des gènes. La caractérisation de la reprogrammation de l'expression des gènes, chez de nombreux allopolyploïdes, révèle une majorité de gènes à expression additive. Le taux des gènes à expression non-additive est variable selon le modèle polyploïde et la méthode d'étude utilisés mais reste largement en dessous de celui des gènes additifs (Pumphrey et al., 2009 ; Akhunova et al., 2010 ; Chague et al., 2010 ; Qi et al., 2012 ; Chelaifa et al., 2013).

En fonction des études, la MPV a été estimée selon deux méthodes principales. Dans la plupart des cas, un mélange équimolaire d'ARN parentaux est réalisé *in vitro* et les proportions des transcrits sont directement mesurées (suivant les différentes technologies) et comparées à l'expression dans les polyploïdes (Wang et al., 2006b; Gaeta et al., 2009; Chague et al., 2010; Qi et al., 2012 ; Chelaifa et al., 2013) (voir mes travaux présentés aux chapitres 3 et 4). Dans d'autres études, la MPV a été mesurée *in silico* en moyennant les valeurs d'expression des progéniteurs (Pumphrey et al., 2009 ; Rapp et al., 2009). Nous avons

néanmoins montré en comparant les deux méthodes dans les blés que la MPV *in silico* n'explique pas toute la variabilité de la MPV *in vitro* estimée à partir des mélanges d'ARN parentaux, sans doute pour des raisons techniques comme par exemple la compétition entre des séquences de gènes similaires à l'hybridation sur les sondes des microarray (Chague et al., 2010).

Basé sur le fait qu'un polyploïde présente un niveau de ploïdie supérieur à celui de ses progéniteurs ou d'un hybride intraspécifique simple, Gianinetti (Gianinetti, 2013) suggère que l'additivité est une somme des valeurs parentales plutôt qu'une moyenne parentale, même si le résultat phénotypique est souvent moyenné du fait des mécanismes de régulation qui contrôlent la traduction du génotype en phénotype. Ainsi, dans le cas de croisement simple (lorsque l'hybride et ses progéniteurs ont le même niveau de ploïdie), l'additivité se caractérise par une contribution de la moitié de chaque parent mais dans un polyploïde l'additivité devrait correspondre dans ce dernier cas à la somme des génomes parentaux (Gianinetti, 2013).

Cependant, les résultats de ma thèse ainsi que ceux d'autre travaux (Pumphrey et al., 2009 ; Akhunova et al., 2010 ; Chague et al., 2010 ; Chelaifa et al., 2013) montrent un taux élevé de gènes exprimés dans les polyploïdes à un niveau égal aux parents et aux moyennes parentales. La dissection de l'expression globale des gènes en celles des homéologues qui les composent, montre également que la majorité de ceux-ci diminuent leur expression dans les polyploïdes comparés aux parents, de façon à maintenir une expression globale (dans l'allopolyploïde) équivalente au niveau d'expression des parents.

L'analyse du modèle blé allohexaploïde rend complexe la validation de l'hypothèse d'additivité de part la présence de trois sous-génomes plutôt que deux. Pour étudier la contribution des génomes parentaux à l'expression globale chez le blé allohexaploïde, la MPV *in silico* a été mesurée suivant les proportions 1 :1 ou 2 :1 pour les génomes AB:D respectivement (Pumphrey et al., 2009). Nos travaux utilisant les sondes PSF et les données RNA-Seq confirment que la MPV appropriée correspond d'avantage à la proportion 2 quantités d'ARN du génome AB pour une quantité d'ARN du génome D (2: 1).

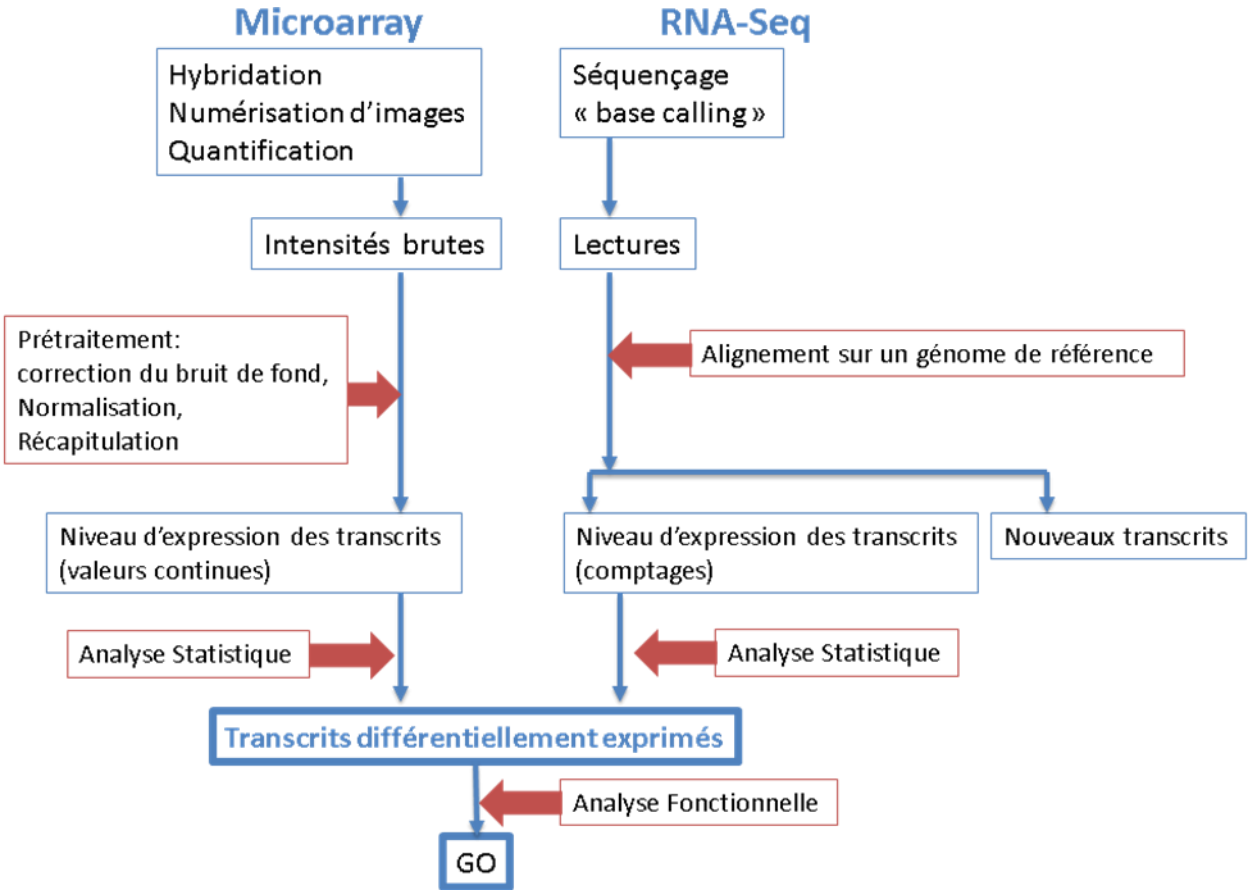


Figure 41:Workflow d'analyses des données microarray et RNA-seq. D'après (Fang et al., 2012).

5.2. Comparaison des technologies Microarray et RNA-Seq pour apprécier l'expression des gènes

5.2.1. Les technologies

Les méthodes microarrays et RNA-Seq ont fait l'objet de nombreuses études comparatives comme outils pour l'analyse de l'expression des gènes (Marguerat et al., 2008 ; Marioni et al., 2008 ; Whitley et al., 2009 ; Matkovich et al., 2010 ; Malone and Oliver, 2011 ; Zhao et al., 2014). Ces deux méthodes présentent des avantages et des inconvénients, ainsi que des similarités (Fig. 41). Bien que récentes, les méthodes RNA-Seq basées sur les NGS sont, très rapidement, devenues des méthodes de choix.

Pour l'analyse des données, il existe trois étapes majeures pour ces deux technologies (Figure 41):

1) le prétraitement des données,
2) l'analyse statistique,
3) et l'analyse fonctionnelle.

Le prétraitement des données microarray inclut une correction du bruit de fond, la normalisation et la synthèse des informations; le prétraitement des données RNA-seq inclut le filtrage des lectures et l'alignement (ou l'assemblage) des lectures. Après le prétraitement des données, le niveau d'expression de chaque transcrit est déterminé.

Pour les microarrays, les niveaux d'expression sont représentés par des nombres continus, alors pour les données RNA-Seq les niveaux d'expression sont représentés par des valeurs discrètes (comptage). L'analyse statistique est, ensuite, effectuée pour identifier les transcrits différentiellement exprimés dans les différentes conditions, les résultats peuvent être analysés au niveau fonctionnel (Fang et al., 2012).

Les technologies microarrays sont plus anciennes que les RNA-Seq avec des temps de processus différents (environ 4 jours contre 10 jours pour les données RNA-Seq d'Illumina). Ils sont dans les deux cas généralement utilisés :

- ✓ pour l'analyse globale de l'expression des gènes,
- ✓ pour déterminer les niveaux d'expressions relatifs
- ✓ ou pour déterminer les changements d'expression des gènes entre différentes conditions.

Les RNASeq permettent d'examiner le transcriptome plus finement, de détecter notamment une expression copie spécifique ou des jonctions d'épissage. Le séquençage haut-débit permet l'identification de nouveaux transcrits. De plus il n'exige pas le séquençage du génome de référence et l'analyse du génome entier avec une résolution à la base près, tandis que l'analyse du transcriptome basées sur la technique microarrays reste limitée à la détection de transcrits connus ou aux organismes dont les génomes ou les gènes ont été séquencés.

La détection, basée sur l'hybridation des ADNc sur les sondes, reste moins sensible et moins spécifique comparée au RNA-seq (Fang et al., 2012). Lorsqu'on étudie des gènes (phylogénétiquement) très proches d'une même famille, et dont les séquences sont très similaires, les microarrays peuvent montrer des hybridations peu spécifiques ou croisées. Cette ambigüité de spécificité de la séquence est corrigée dans les RNA-seq, même si une lecture peut s'aligner sur plusieurs séquences identiques du génome, ce qui nécessite de déterminer avec précision ces séquences de références comme je l'ai développé dans les résultats présentés dans le chapitre 4. Le séquençage des deux extrémités (en paired-end) augmente la précision dans l'alignement permettant ainisi de contourner le problème de taille de lecture.

Toutefois, il existe un nombre de biais spécifiques au séquençage RNA-Seq haut-débit, à savoir:

- lors de la préparation de la bibliothèque d'ARNm, la fragmentation de l'ARN par 'sonication' ou digestion enzymatique peut induire un biais dans la composition et représentation de la séquence (Roberts et al., 2011);

- les plateformes de RNA-Sequencing en NGS utilisent toutes une transcription reverse suivie d'une amplification par PCR avant le séquençage. L'ajout d'amorces (hexamères) aléatoires sur l'ADNc résulte en un biais dans la composition en nucléotide des lectures (en début de séquence). Ce biais influence le positionnement et par conséquent la distribution des lectures le long du transcrit exprimé (Hansen et al., 2010). La PCR pourrait également introduire des biais car elle est influencée par le contenu en GC ou par la longueur de la séquence matrice produisant du fait une amplification non-linéaire (Dohm et al., 2008; Risso et al., 2011).

De nombreux outils ont été développés afin de corriger ces biais (Hansen et al., 2010 ; Jones et al., 2012).

Pour une étude du transcriptome, ces deux techniques (RNAseq et microarrays) présentent des résultats concordants avec des particularités pour chacune permettant une complémentarité (Bloom et al., 2009 ; Agarwal et al., 2010 ; Liu et al., 2010 ; Malone and Oliver, 2011).

La détection des gènes faiblement exprimés restent problématique pour les deux techniques, mais le séquençage RNA-Seq reste mieux adapté pour détecter les transcrits faiblement exprimés et identifier des variants d'épissages. Pour conclure sur les deux techniques, la comparaison des méthodes de microarrays et RNAseq sur un même matériel biologique montre une forte corrélation entre les profils d'expression de gènes générés par les deux plateformes (Whitley et al., 2009), ce que j'ai également observé durant mes travaux de thèse.

5.2.2. Adéquation des technologies microarray et RNASeq pour l'analyse de l'expression des gènes dupliqués chez le blé.

Les résultats d'analyse de l'expression globale des gènes par la méthode microarray (au chapitre 2), ou l'outil sonde parent spécifique (au chapitre 3) ou encore par la technologie RNA-Seq (chapitre 4) convergent dans la même direction. Pour les évaluer, il convient de comparer les principaux résultats obtenus avec ces trois méthodes dans les principales comparaisons réalisées dans mes travaux de thèse:

Comparaison entre les allohexaploïdes naturels et synthétiques

Les données microarray montrent une très forte similarité dans l'expression globale des gènes entre le blé naturel allohexaploïde cv Courtot et son allotétraploïde extrait « Tetra-Courtot » (98.6% des gènes étudiés ont le même niveau d'expression). Les sondes parents spécifiques (« PSF ») montrent aussi une majorité de gènes avec une expression globale égale entre ces deux allopolyploïdes (92.6% de concordance entre Courtot vs Tetra-Courtot). Les données RNA-Seq montrent moins de similarité entre le blé naturel allohexaploïde cv Courtot et son allotétraploïde « Tetra-Courtot », avec seulement 68.2% des gènes ayant une expression globale égale. Pour cette comparaison entre ces deux allopolyploïdes, l'outil microarray utilisé montre, pour les 1,4% des gènes différentiellement exprimés, une majorité de gènes sur-exprimés (91.9%) dans le blé hexaploïde Courtot alors que l'outil PSF en révèle 89.5%. Pour les 37,8% des gènes différentiellement exprimés, l'outil RNA-Seq ne montre pas une différence aussi importante entre les gènes sur- ou sous-exprimés dans le blé naturel allohexaploïde comparé à son allotétraploïde extrait.

Comparaisons de l'allohexaploïde naturel cv Courtot au diploïde Ae. tauschii

Les comparaisons montrent:

- avec la méthode affymetrix microarray où *Ae. tauschii* est une des accessions AttD54 ou AtsD36, que ~92% de gènes ont le même niveau d'expression,
- avec la méthode PSF et les mêmes accessions d'*Ae. tauschii* (AttD54 ou AtsD36), que 73.5% de gènes montrent une expression globale égale,
- avec l'outil RNA-Seq où une troisième accession d'*Ae. tauschii* (T109) est utilisée, on observe 55.1% de gènes avec un niveau d'expression globale égale.

Dans ces 3 méthodes, les gènes différentiellement exprimés sont majoritairement surexprimés dans le blé naturel allohexaploïde cv Courtot (en considérant uniquement les gènes différentiellement exprimés, la méthode microarray montre 79.9% de gènes sur-exprimés dans l'allohexaploïde, la méthode PSF 73.5% et 59.5% pour la méthode RNA-Seq).

Comparaisons entre l'allotétraploïde Tetra-Courtot et le diploïde Ae. tauschii

Les comparaisons montrent:

- avec la méthode microarray, ~86% de gènes à une expression globale égale (avec les accessions AttD54 ou AtsD36) ;
- avec la méthode PSF et la méthode RNA-Seq, que l'on obtient 50.2% et 47.6% (respectivement) de gènes présentant un niveau d'expression similaire. Ces comparaisons entre l'allotétraploïde Tetra-Courtot et le diploïde *Ae. tauschii* montrent beaucoup plus de différence entre l'outil microarray et les techniques plus précises, disséquant l'expression globale des gènes. Ainsi, il est possible d'atteindre une qualité d'analyse de l'ordre du sous-génome ou des homéologues.

Interprétation

Les différences de résultats obtenus avec l'outil microarray, l'outil PSF, et l'outil RNA-Seq peuvent s'interpréter par de multiples raisons d'ordre techniques ainsi que par le dispositif expérimental et les méthodes d'analyses statistiques:

(i) Les sondes dans les puces microarray peuvent correspondre à un domaine commun d'une famille de gènes. Dans ce cas, plusieurs gènes différents d'une même famille peuvent s'hybrider. Cet exemple, combiné au fait qu'une sonde peut par définition être saturée par un excès d'ARN s'hybridant, peut sur-représenter le nombre de gènes non différentiellement exprimés. Dans l'étude de l'expression des sous-génomes, cette méthode est par conséquent moins spécifique et moins précise que la méthode PSF qui utilise des sondes spécifiques des parents (donc hors du domaine commun et conservé), et moins sujet à des hybridations croisées avec des ARNm de gènes de mêmes familles.

(ii) La technique RNASeq demeure la plus spécifique des trois, car en théorie la méthode d'alignement des lectures que j'ai utilisé dans ma thèse, sélectionne uniquement les lectures s'alignant de façon unique avec un « match » à 100% d'identité des séquences et spécifique d'un seul homéologue. D'ailleurs, on note que la méthode

des PSF présentent des résultats très similaires à la méthode RNA-Seq pour la comparaison des génotypes allotétraploïde et diploïde, car les sondes sont également spécifique des sous-génomes.

(iii)Une autre différence entre ces techniques, avec les méthodes RNA-Seq trois réplicats biologiques assez homogènes ont été utilisées et la variance globale est donc assez faible augmentant ainsi la détection de gènes différentiellement exprimés. Avec la méthode microarray seulement deux réplicats biologiques ont été utilisés, ce qui augment la variance et diminue alors la détection des gènes différentiellement exprimés.

Les PSF et les RNA-Seq montrent de nombreux résultats similaires. Ces deux méthodes montrent un pic de distribution à 2/3 pour le ratio d'expression des homéoallèles AB dans l'allohexaploïde naturel comparé à leur expression dans l'allotétraploïde « Tetra-Courtot » et un pic à 1/3 pour le ratio d'expression des homéoallèles D dans l'allohexaploïde comparé à leur expression dans le diploïde. Néanmoins, la méthode PSF reste moins résolutive car il n'a pas été possible de développer des PSF pour un grand nombre de gènes analysés, ni de séparer les homéologues A_h et B_h. La méthode RNASeq reste la plus performante et la mieux adaptée pour étudier de façon spécifique l'expression des homéologues de la plupart des gènes, la spécificité étant basée sur les différences des séquences entre ces copies dupliquées.

5.3. Conclusions sur les réponses aux changements du niveau de ploïdie chez le blé

Plusieurs études ont montré que la polyploïdie induit des changements dans l'expression des gènes sur le cours terme (rapide) ainsi que sur le long terme (évolutionnaire) (Wendel, 2000 ; Adams et al., 2003 ; Hegarty et al., 2006 ; Wang et al., 2006b ; Chaudhary et al., 2009); (Pumphrey et al., 2009 ; Rapp et al., 2009 ; Akhunova et al., 2010 ; Chague et al., 2010 ; Buggs et al., 2011a ; Grover et al., 2012 ; Li et al., 2014 ; Rambani et al., 2014; Zhang et al., 2014). Ces changements représenteraient des réponses adaptatives de la nouvelle espèce polyploïde (Paterson et al., 2011 ; Rambani et al., 2014 ; Xu et al., 2014).

Différents mécanismes interviennent dans la régulation de l'expression des gènes et le devenir des homéologues (Comai, 2005 ; Chen and Ni, 2006) comme : la délétion de gènes, les interactions incompatibles ou modifiées entre les voies de régulations, les changements de dosage de gènes, le partitionnement de l'expression (Chalhoub et al., 2014), la compensation de l'expression ou encore la néo- et sous-fonctionnalisation (Comai, 2000 ; Birchler et al., 2003 ; Osborn et al., 2003; Riddle and Birchler, 2003 ; Adams and Wendel, 2005b ; Hovav et al., 2008 ; Chaudhary et al., 2009 ; Ha et al., 2009a ; Pang et al., 2009).

Les études d'expression de gènes dans différentes espèces allopolyploïdes montrent que si l'expression des gènes dans l'allopolyploïde est une simple moyenne de celle des parents pour une bonne partie des gènes, une autre partie significative, plus ou moins importante selon le polyploïde et l'étude, montrent aussi une expression non-additive (Hegarty et al., 2006 ; Tate et al., 2006 ; Wang et al., 2006a ; Gaeta et al., 2009; Chague et al., 2010 ; Flagel and Wendel, 2010 ; Qi et al., 2012 ; Gianinetti, 2013). Les travaux menés dans notre équipe, montrent que la proportion des gènes additifs et non-additifs varie en fonction du génotype des parents, par conséquent la stabilité des allopolyploïdes dépend de cette diversité. Chagué et al. (Chague et al., 2010) utilisent un allohexaploïde synthétique, dont le donneur du génome AB est l'espèce naturelle tétraploïde *T. turgidum* spp. *durum* cv Joyau (Mestiri et al., 2010), et trouvent un nombre relativement important de gènes non-additifs (~2000 gènes) (Chague et al., 2010). Dans mes travaux de thèse présentés au chapitre 2 dans l'article (Chelaifa et al., 2013), j'utilise des allohexaploïdes synthétiques plus stables du point de vue de la régularité de la méiose, ayant les même parents donneur du génome D

que ceux dans Chagué et al. (2010) mais le donneur du génome AB a été remplacé par l'allotétraploïde extrait « Tetra-Courtot ». Dans (Chelaifa et al., 2013), je trouve que l'expression globale des gènes montrent une quasi-totale additivité (>99%). Ces résultats suggèrent que l'allohexaploïdisation n'induit pas des changements importants dans l'expression des gènes si les donneurs des génomes AB et D sont similaires à ceux du blé naturel allohexaploïde. C'est la raison pour laquelle, j'ai utilisé pour l'analyse de l'expression basée sur les RNA-Seq, l'allohexaploïde synthétique TC109 le plus stable et le plus proche de l'allohexaploïde naturel, caractérisé précédemment dans le laboratoire (Mestiri et al., 2010).

Cette additivité quasi-totale suggère un équilibre de dosage (Birchler et al., 2005 ; Veitia et al., 2008), et une concertation entre les copies homéologues pour la contribution à l'expression (Adams et al., 2003 ; Adams and Wendel, 2005b; Jackson and Chen, 2010).

Ainsi, nous avons pu mieux comprendre la régulation de l'expression des gènes à différents niveaux de ploïdie en disséquant l'expression globale plus finement au niveau du partitionnement entre les différents homéologues.

Ceci a été rendu possible par les avancées récentes du séquençage du génome du blé et la disposition des séquences des homéologues A_h, B_h et D_h de 8605 gènes du blé allohexaploïde naturel (IWGSC, 2014). Aux résultats du séquençage du génome de blé allohexaploïde, nous avons combiné notre système original, nous permettant d'interpréter les réponses obtenues suite à la diminution suivie de la ré-augmentation du niveau de ploïdie (Chague et al., 2010; Mestiri et al., 2010 ; Chelaifa et al., 2013). Ainsi mes travaux révèlent de nombreuses particularités de la régulation de l'expression des gènes et de son partitionnement entre les différents homéologues:

(i) Le partitionnement de l'expression des homéologues est clairement établi dans le blé naturel allohexaploïde, son allotétraploïde extrait « Tetra-Courtot » et le blé allohexaploïde nouvellement synthétisé. La majorité des homéologues contribuent ainsi à l'expression globale des gènes, de manière équivalente. D'autres homéologues montrent un biais d'expression spécifique vers un des sous-génomes, sans être significativement plus dirigé vers une dominance d'un des sous-génomes (IWGSC, 2014; Pfeifer et al., 2014). Cette absence de dominance d'un sous-génome a été révélée aussi dans le génome du colza, un autre allopolyploïde récent (Chalhoub et al., 2014) mais également dans certains

génomes d'anciens polyploïdes (Garsmeur et al., 2014). Ceci contraste avec la dominance, d'un sous-génome sur un autre, observée dans d'autres polyploïdes anciens ou récents (Buggs et al., 2011b; Schnable et al., 2011; Ilut et al., 2012; Tang et al., 2012; Yoo et al., 2013).

(ii) Mes travaux suggèrent qu'une concertation dans le partitionnement et le niveau d'expression des homéologues est établie dans le blé. Ainsi la majorité des homéologues ont leur expression qui augmente lorsqu'ils sont séparés et qui diminue lorsqu'ils sont rassemblés dans un niveau de ploïdie supérieur. L'expression des homéologues A_h et B_h dans le blé allohexaploïde naturel ou synthétique est généralement égale à 2/3 de leur niveau d'expression dans l'allotétraploïde extrait « Tetra-Courtot » ; tandis que celle de l'homéoallèle D_h est au 1/3 du niveau d'expression mesuré dans le diploïde *Ae. tauschii*.

L'expression des homéoallèles A_h et B_h dans l'allotétraploïde extrait est à son tour généralement plus faible que celle de D_t dans l'espèce diploïde *Ae. tauschii*.

Les niveaux d'expression de ces gènes dupliqués seraient, pour la majorité, inversement proportionnels au niveau de ploïdie.

Ainsi pour la plupart des gènes (~94.4%) l'expression des homéologues augmente quand on sépare le génome B_hA_h du génome D_h (dans l'allotetraploïde extrait), puis diminue quand on remet ensemble les génomes B_hA_h et les génomes D_t dans l'allohexaploïde synthétique. De plus, les niveaux d'expressions sont égaux entre allohexaploïde naturel et synthétique. Cette concertation semble donc agir sur l'augmentation et la diminution de l'expression des homéologues de façon à maintenir l'expression globale du gène à un niveau adéquat. Ceci pourrait représenter un mécanisme de rétention des copies des gènes dupliqués pour conserver la fonction (Coate, 2013 ; Keane et al., 2014).

Ainsi, mes résultats suivent l'hypothèse d'un équilibre de dosage ou « dosage-balance hypothesis » entre les homéologues ou les gènes dupliqués (Veitia et al., 2008 ; Birchler and Veitia, 2012). Cette hypothèse prédit que, dans un réseau, les gènes présentant plusieurs interactions clés avec d'autres composants du métabolisme sont préférentiellement conservés car leur élimination perturberait la stœchiométrie de nombreuses interactions (Mach, 2011).

On peut se poser la question si la régulation de l'expression des homéologues ne serait pas du au fait que dans mes travaux ainsi que dans d'autres études (Rambani et al., 2014 ; Xu et al., 2014), une quantité égale d'ARN est comparée quelque soit le niveau de ploïdie?

En effet, le volume de la cellule augmente généralement avec la taille du génome (Otto, 2007), bien que la relation exacte entre niveau de ploïdie et le volume de la cellule varie en fonction de l'environnement et de la taxonomie des plantes. La ploïdie n'affecte pas la densité des cellules pour un tissu équivalent, mais affecte le conditionnement de la cellule: généralement observable par le volume des cellules deux fois plus grandes et contenant deux fois plus d'ADN. Ainsi, une feuille issue d'une espèce de blé diploïde, allotétraploïde ou allohexaploïde contiendrait autant de cellules mais la taille de ces feuilles augmente entre les espèces diploïdes et allopolyploïdes. On ignore si la quantité d'ARN transcrit est identique entre les différents niveaux de ploïdie ou si ce niveau augmente (ou diminue) avec le niveau de ploïdie.

Néanmoins, l'adéquation d'utiliser la normalisation RPKM pour comparer le transcriptome à partir de quantités d'ARNs égales quelque soit le niveau de ploïdie (avec un niveau d'expression des homéologues qui diminue inversement avec le niveau de ploïdie), viendrait du fait qu'on aligne sur les 8605 homéologues une proportion équivalente d'environ ~10% de lectures brutes (RNA-Seq) (Table S1 de l'article présenté dans le chapitre 4):

(i) issus d'*Ae.tauschii* (diploïde), théoriquement un alignement sur les 8605 séquences homéologues D seulement ;

(ii) issus de l'allotetraploïde extrait, théoriquement un alignement sur les 17 210 séquences homéologues A_h et B_h;

(iii) issus des allohexaploïdes, théoriquement un alignement sur les 25 815 séquences des trois homéologues (A_h, B_h et D_h).

Si la proportion de ~10% de lectures brutes RNA-Seq alignées sur les espèces de trois niveaux de ploidie n'étaient pas similaires, la normalisation RPKM ne seraient pas adéquate pour comparer l'expression des homéologues entre les trois niveaux de ploïdie. Une normalisation alternative reflétant mieux le nombre de lectures alignables par million de lectures brutes serait plus approporiée dans le cas où les proportions ne sont pas les mêmes.
D'autre part, la normalisation utilisée (par million de lecture) rendrait l'expression relative d'un gène comparable même si les conditions comparées ne produisaient pas la même quantité d'ARN par cellule au départ.

Par ailleurs, l'hypothèse de dosage balance, utilisée ci-dessus, tendrait plutôt vers un contrôle de la quantité d'ARN total transcrit par gène pour maintenir les rapports stœchiométriques nécessaires aux réseaux de régulations de gènes. Sous l'effet de l'augmentation du niveau de ploïdie, les gènes dupliqués n'augmenteraient pas l'activité globale de leur fonction pour maintenir une régulation solide. Cette aptitude pourrait être à l'origine d'une évolution vers une diversification de leur fonction (Keane et al., 2014). La robustesse de la régulation est également une force pour l'évolution des gènes dupliqués, notamment pour s'adapter aux différentes conditions environnementales (Keane et al., 2014). A l'échelle de la cellule, un niveau de ploïdie supérieur pourrait jouer un rôle dans l'acquisition de nouvelles fonctions permettant à la plante d'acquérir un effet « hétérosis » tel que chez les hybrides.

Il n'est pas possible de déterminer si le génome B_hA_h du blé naturel allohexaploïde a été présélectionné lors de l'allohexaploïdisation naturelle, menant vers un blé allohexaploïde très stable, ou si cette stabilité résulte de modifications qui ont eu lieu pendant sa co-résidence, pour près de 10 000 ans, avec le génome D, induisant des possibles modifications 'adaptatives'. La forte proportion de gènes à expression additive dans les polyploïdes synthétique ayant ce génome B_hA_h ainsi que la forte proportion de gènes à expression globale égale entre les allohexaploïdes naturels et synthétiques 98.36% (Chelaifa et al., 2013) suggère un lien avec la forte stabilité de ces allohexaploïdes.

Récemment, il a été suggéré que les changements majoritaires entre le blé allotétraploïde extrait et différentes sous-espèces de blé allotétraploïde *T. turgidum*, au niveau de l'expression globale utilisant l'outil microarray, seraient dus à l'acquisition ou la modification de certaines voies de régulation de l'expression dans le génome B_hA_h, qui sont absentes dans les allotétraploïdes naturels (Zhang et al., 2014). Ceci aurait pu se dérouler lors de sa co-résidence avec le génome D dans le blé allohexaploïde selon les auteurs (Zhang et al., 2014). Nos travaux utilisant les outils micro-arrays ou RNA-Seq, comparant deux allotétraploïdes extraits différents du tétraploïde naturel *T. turgidum* spp. *durum* cv Joyau, confirment le biais observé par (Zhang et al., 2014) d'une plus forte proportion de gènes plus exprimés dans les allotetraploïdes extraits que dans les allotétraploïdes naturels. Toutefois, les comparaisons montrent aussi une prépondérance des homéologes D_h surexprimés dans le blé hexaploïde naturel par rapport aux homéologues D_t dans les allohexaploïdes synthétiques (Chapitres 3 et 4). Il est à souligner que les outils microarray et RNA-Seq utilisent des

séquences références des gènes et des homéologues du blé allohexaploïde naturel. En reprenant ces informations et les résultats obtenus lors de ma thèse, il est tentant de suggérer un biais technique due à une plus forte détection de l'expression des homéologues de références du blé hexaploïde naturel et son tétraploide extrait comparés à leur équivalent dans les blés des espèces tétraploïdes et diploïde apparentées.

Mes analyses de l'expression des homéologues basées sur l'outil RNA-Seq montrent que leurs niveaux d'expression sont coordonnés et concertés dans les différents niveaux de ploïdie. Ces conclusions étant basées sur les gènes qui possèdent encore les trois homéologues dans le blé hexaploïde. Pour compléter ces travaux et vérifier si les changements d'expression sont dus à des modifications génétiques (Akhunova et al., 2010 ; Saintenac, 2011 ; Brenchley et al., 2012 ; Akhunov et al., 2013 ; Pont et al., 2013) et/ou épigénétiques (Shaked et al., 2001 ; Zhao et al., 2011), il serait intéressant d'étendre l'analyse sur l'ensemble des gènes identifiés dans le blé comme singletons (en copie unique) ou en deux copies seulement.

Ayant eu les séquences des 8605 gènes en triplets d'homéologues à la publication de la séquence 'brouillon' du génome du blé hexaploïde (IWGSC, 2014), c'est-à-dire seulement au mois d'août dernier, je n'ai pas réalisé l'analyse des fonctions des différentes catégories d'expression identifiées. Cette analyse demeure importante et devrait se poursuivre, afin de nous révéler les fonctions et termes de gene ontology (GO) activés ou reprimés en réponse à la polyploïdie.

Annexes

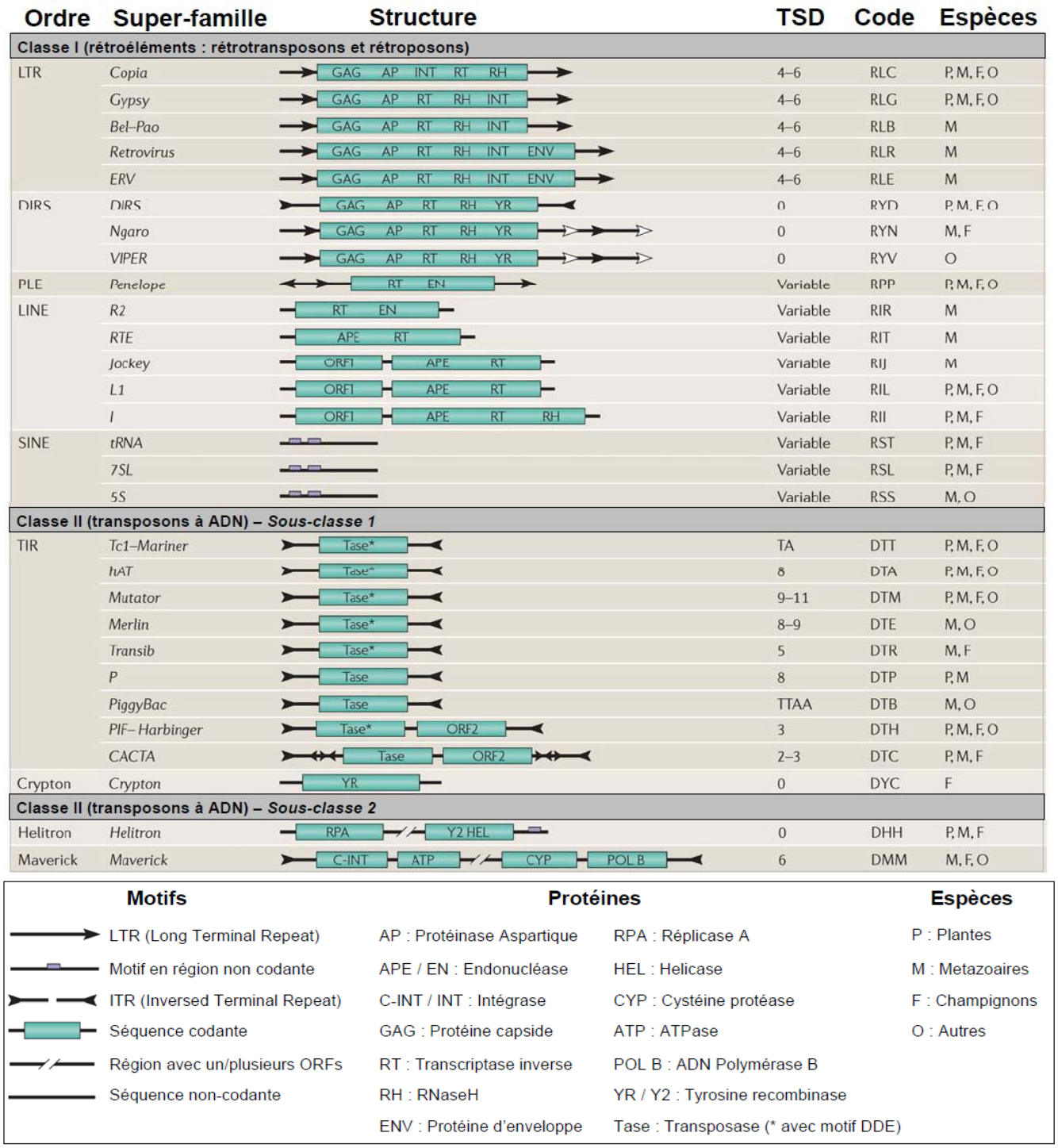


Figure Annexe 1: Classification des différents éléments transposables sur 5 niveaux: classe, sous-classe, ordre, super-famille, famille d'après (Wicker et al., 2007).

Le dernier niveau, sous-famille, n'est pas illustré sur cette figure. Les éléments caractéristiques de chaque catégorie sont mentionnés, ainsi que les espèces dans lesquelles ils ont été trouvés. La colonne TSD indique la taille de ceux-ci pour les éléments de cette famille. Code indique le préfixe à faire figurer devant l'élément pour sa nomenclature. ORF1 et ORF2 représentent des protéines dont la fonction est inconnue.

Annexe 1: Les éléments transposables dans les génomes du blé

Cette partie sur les éléments transposables reprend, pour la majorité, les thèses précédentes réalisées au sein de mon équipe (Charles, 2010; Belcram, 2014).

1. La classification des TEs

Plusieurs classifications se sont succédées, la première reposant sur le mécanisme de transposition (Finnegan, 1990 ; Capy, 1998).

Les TEs se répartissent en deux grandes classes :

- les éléments de classe I utilisent un intermédiaire ARN pour transposer. Suivant un mécanisme de 'copier-coller', les éléments de classe I se copient et s'insèrent à une autre place dans le génome en utilisant un intermédiaire ARN ;
- les éléments de classe II codent une enzyme « transposase » qui reconnaît les extrémités de l'élément, l'excise et l'insère ailleurs dans le génome par « couper-coller » (Wicker et al., 2007).

Les TEs de classe I sont dits rétroéléments, par référence à la transcriptase inverse leur permettant de copier l'ARN en ADN. Il existe deux types de rétroéléments : les rétrotransposons et les rétroposons.

Les rétrotransposons sont reconnaissables par leurs LTR (Long terminal Repeat) qui correspondent à des longues séquences répétées au début et à la fin de l'élément.

Les éléments de classe II sont appelés transposons à ADN.

D'autres classifications ont été proposées : en 2007, le modèle hiérarchique (Figure Annexe 1, (Wicker et al., 2007)) classent les TEs d'eucaryotes selon leur mode de transposition (au niveau mécanistique et enzymatique). Cette classification prend en compte d'autres modèles (Jurka et al., 2005) et propose des règles pratiques, notamment la règle dite « 80-80-80 » basée sur la similarité de séquences : deux éléments appartiennent à la même famille s'ils ont plus de 80% d'identité sur 80% de leur séquence sur un minimum de 80 pb.

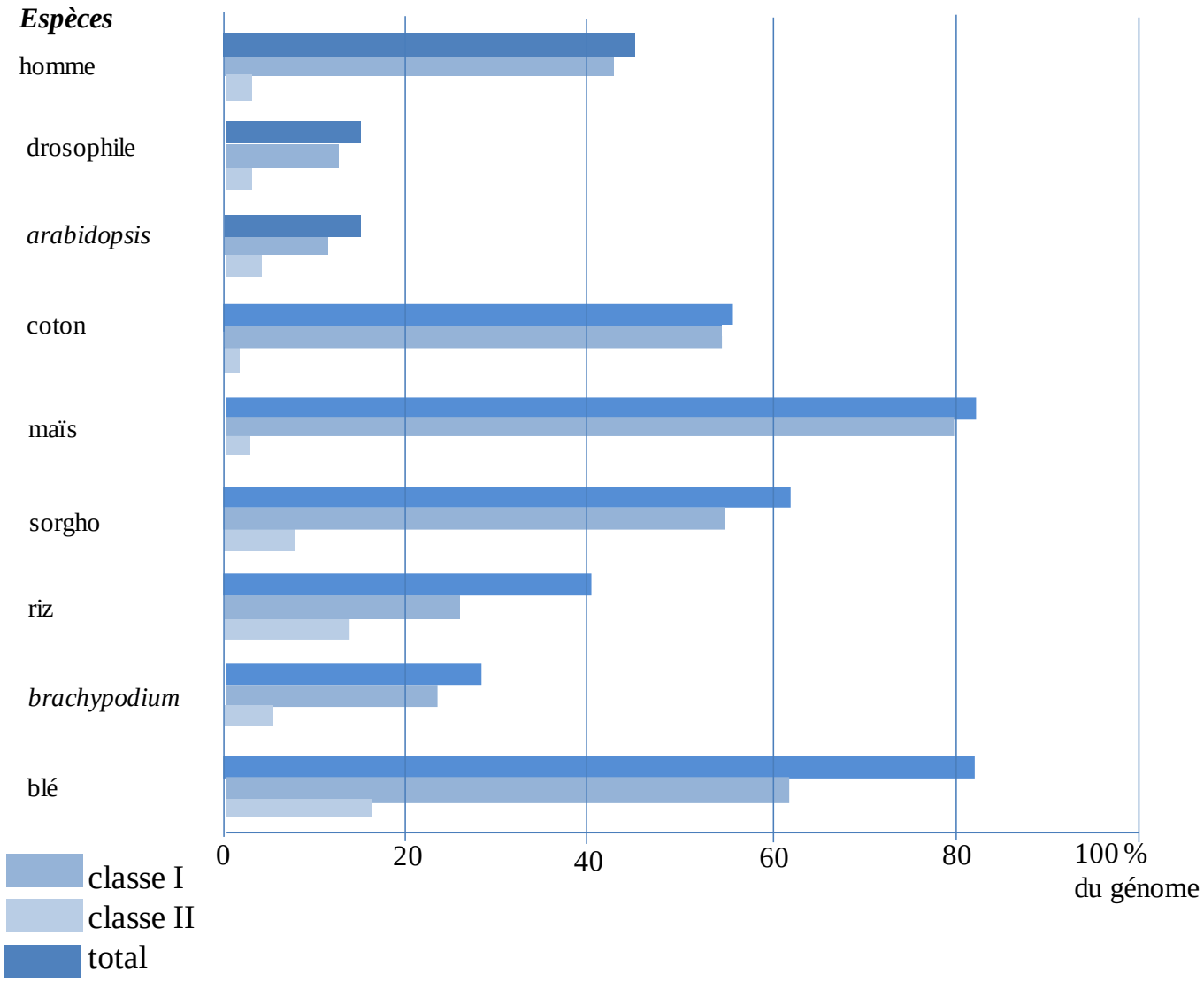


Figure Annexe 2: Proportions des TEs de classe I et II dans différentes espèces animales et végétales.

Ces proportions viennent des données de séquencage (Initiative, 2000 ; Lander et al., 2001 ; Kaminker et al., 2002 ; Project., 2005 ; Paterson and al., 2009)ou d'études représentatives (Hawkins et al., 2006; Paux et al., 2006; Piegu et al., 2006; Charles et al., 2008). Ces données ont ensuite été complétées par les données d'une revue récente des TEs dans les eucaryotes (Pritham, 2009), en particulier pour les proportions de classe I et II.

Le système classe I / classe II a été conservé et enrichi de 5 niveaux : sous-classe, ordre, super-famille, famille et sous-famille.

- Les classes restent définies par la présence ou non d'un intermédiaire ARN lors de la transposition (transposons à ARN ou transposons à ADN).

- Les sous-classes séparent les transpositions par 'copier-coller' des transpositions par 'couper-coller'. Tous les éléments de classe I appartiennent donc à la même sous-classe ('copier-coller'), alors que les éléments de classe II peuvent appartenir à l'une ou à l'autre.

- Les ordres séparent les transpositions ayant des caractéristiques enzymatiques et organisationnelles différentes.

- Les super-familles différencient des éléments ayant une même stratégie de réplication, mais une conservation au niveau protéique très limitée (Figure Annexe 1, super-familles des copia et des gypsy).

- Les familles sont composées d'éléments présentant une forte conservation au niveau protéique (>80%). Certaines familles ont des membres se regroupant en sous-familles sur des critères de similarité (conservation nucléique ou regroupement dans des arbres phylogénétiques). Exemple : Wis et Angela sont deux sous-familles de BARE-1.

2. L'importance des TEs

La proportion en TEs est très variable d'une espèce à une autre : chez les plantes dicotylédones, les TEs représentent 57% du génome du coton *G. raimondii* (Wang et al., 2012) en comparaison des 14% du génome d'*A.thaliana* (Initiative, 2000). Chez le règne animal, ils représentent 45% du génome chez l'homme (Lander et al., 2001) et seulement 15% du génome de la drosophile (Kaminker et al., 2002). Cette variation en TEs est considérable dans la famille des *Poaceae* (Figure Annexe 2), variant de 23% chez brachypodium à 82% chez le blé (Paux et al., 2006 ; Charles et al., 2008).

Les TEs sont des éléments très importants et dynamiques des génomes du blé. Il existe des variations de taille très importantes entre les différentes espèces de blé pouvant atteindre plusieurs centaines de Mb pour des espèces ayant le même niveau de ploïdie (Bennett and Smith, 1976 ; Bennett and Smith, 1991).

La proportion des TEs de classes I et II est également très variable selon les espèces (Pritham, 2009) (Figure Annexe 2). Les éléments de classe I représentent plus de 60% des génomes du blé, plus précisément 66% du chromosome 3B du blé (Choulet et al., 2014) mais seulement 10% du génome de la drosophile. Le séquençage du génome du blé révèle une différence dans la distribution des TEs dans les sous-génomes A, B, D (IWGSC, 2014) : les TEs de classe I (les rétroéléments) sont plus abondants dans le sous-génome A comparé aux sous-génomes B et D (A>B>D). Les éléments de classe II sont près de 10 fois plus importants dans le génome du blé (20%) que celui du maïs (3.2%) (Schnable et al., 2009). Les travaux récents du séquençage du chromosome 3B montrent que 60% de la séquence est composée de TEs de classe II, qui sont majoritairement des éléments CACTA (Choulet et al., 2014).

La taille des TEs peut également varier de quelques dizaines de paires de bases (pb) à quelques dizaines de millier de paires de bases (kb), selon la famille des éléments. Certaines espèces contiennent plusieurs centaines de familles d'éléments alors qu'une seule famille d'éléments peut représenter la grande majorité des TEs présents dans un génome (élément *Alu* chez l'homme, *BARE* chez l'orge).

3. La dynamique des TEs

3.1. La prolifération des TEs

Les différences de proportion des TEs entre les diverses espèces proviennent de variations dans la dynamique de prolifération. La prolifération des TEs résulte de deux forces d'évolution antagonistes : l'insertion et l'élimination (SanMiguel et al., 1996 ; Bennetzen and Kellogg, 1997 ; Petrov et al., 2000 ; Kidwell, 2002 ; Wendel et al., 2002 ; Bennetzen et al., 2005 ; Hawkins et al., 2006 ; Piegu et al., 2006 ; Zuccolo et al., 2007). Ces deux forces sont difficilement séparables lors de l'étude de la dynamique globale des TEs dans un génome, dans le sens où des pics apparents d'insertions de TEs dans un génome peuvent résulter d'une forte activité d'insertion et/ou d'une faible vitesse d'élimination.

La disponibilité et l'abondance des séquences de TEs à l'échelle génomique ont permis de caractériser ces deux forces (Wicker et al., 2003; Wicker et al., 2007); (Gao et al., 2004); (Piegu et al., 2006) afin de déterminer :

- la proportion des TEs et les différentes familles dans le génome;
- la proportion de copies complètes, ou tronquées;
- la distribution des TEs le long des chromosomes;
- l'estimation des dates d'insertion des rétrotransposons ayant leur deux LTR.

La prolifération des TEs n'est pas constante au cours de l'évolution des espèces, ni homogène le long des chromosomes. On distingue des périodes d'activité très fortes et des périodes de plus faible intensité. La plupart des TEs, qui forment le chromosome 3B se sont insérés avant la polyploïdisation (il y a environ 0.5 millions d'années (Ma)), et sont devenus moins actif par la suite (Choulet et al., 2014). Les événements récents d'insertions de TEs sont homogènes dans les régions distales (R1 et R3) et proximale (R2) du chromosome 3B (Figure Annexe 3). Les insertions plus anciennes (>1.5 Ma) ont été plus rapidement éliminées dans les régions distales du chromosome 3B (Choulet et al., 2014). La densité de distribution des TEs le long du chromosome 3B n'est pas aléatoire, il y a une plus faible densité dans les régions distales R1 (73%) et R3 (68%), et une plus forte densité dans la région proximale R2 (88%). La région centromérique-péricentromérique montre une plus forte densité de TEs (93%) (Choulet et al., 2014). A l'échelle du génome entier, sous l'hypothèse d'une dynamique d'insertion/élimination similaire entre les 3 sous-génomes A, B et D, la distribution en rétrotransposons est plus faible dans le sous-génome B (IWGSC, 2014).

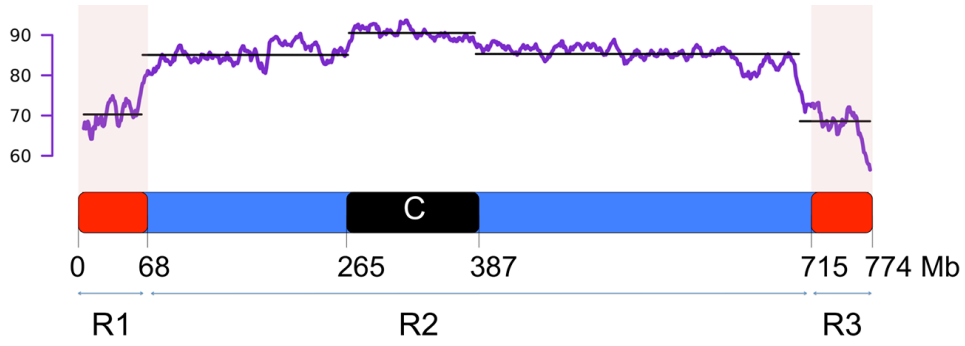


Figure Annexe 3: Distribution et ségmentation des TEs le long du chromosome 3B du blé.

D'après (Choulet et al., 2014) : Le gaphe du dessus décrit le contenu en TEs le long du chromosome 3B du blé. Les régions distales R1 et R3 sont représentées en rouge, la région centromérique/péricentromérique est représentée par le C.

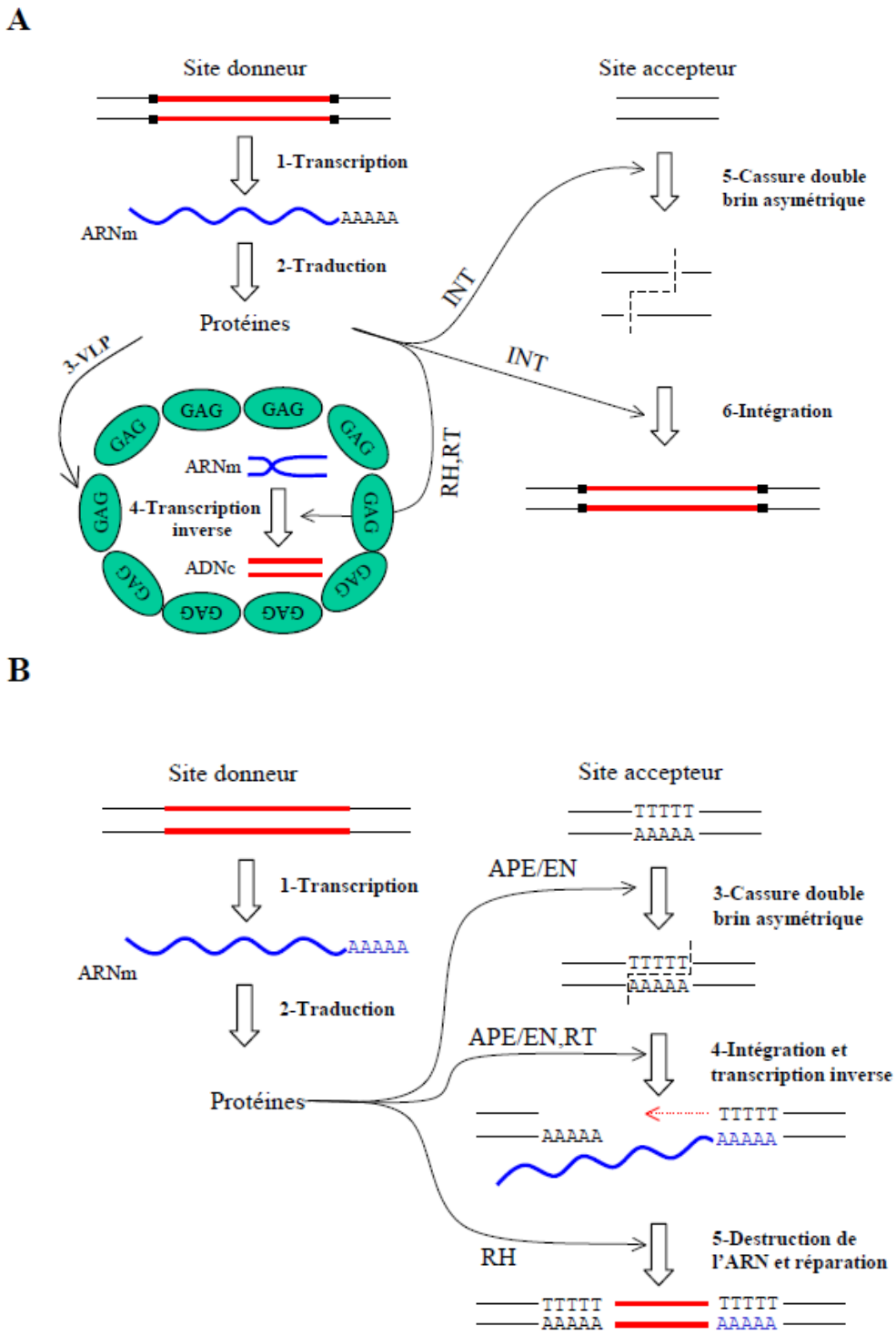


Figure Annexe 4:Mécanismes de transposition des principaux TEs de classe I, d'après (Sabot et al., 2004).

(A) Transposition des rétrotransposons. Les carrés noirs indiquent les TSDs. (B) Transposition des rétroposons. Les différentes étapes des transpositions sont détaillées dans le texte correspondant.

De plus, elle semble différente d'une espèce à une autre et d'une famille de TEs à une autre. Par exemple, dans le chromosome 3B du blé, quelques transposons à ADN (ou TEs de classe II : Mutator, Harbinger et MITE) sont répérés très proches des gènes, tandis que les rétrotransposons (TEs de classe I) et les transposons CACTA (TEs de classe II) sont très distants des gènes (Choulet et al., 2014 ; IWGSC, 2014). Chez les angiospermes, les estimations des dates d'insertion des rétrotransposons n'excèdent pas 3 Ma alors que chez les gymnospermes, la plupart des estimations sont supérieures à 35 Ma (Nobuta et al., 2008).

3.2. Mécanismes d'insertion des TEs

De nombreux processus associés à l'allopolyploïdie (redondance génétique, « choc génomique »…) induiraient une accumulation des insertions des TEs. Cette mobilité des TEs serait une réponse à un stimulus spécifique (Grandbastien et al., 2005).

Les TEs sont dits autonomes lorsqu'ils peuvent coder les protéines essentielles à leur transposition. Ces TEs ont, toutefois, besoin de la machinerie de la cellule hôte pour traduire ces protéines. Lorsqu'un élément autonome est actif, il peut transposer mais aussi induire la transposition d'éléments de la même famille, qui sont non-autonomes (ne codant pas toutes les protéines essentielles à leur transposition). Un élément non-autonome peut avoir conservé les sites de reconnaissance (peu spécifiques) de la machinerie de transposition.

Les mécanismes de réplication des TEs de classes I et II sont très différents. Les éléments de classe I passent par un intermédiaire ARN. Le mode de transposition par 'copier-coller' des éléments de classe I les rend potentiellement très invasifs pour un génome. Il n'est donc pas surprenant de les voir représenter parfois une grande partie d'un génome. C'est particulièrement le cas des rétrotransposons, plus fréquents chez les plantes que chez les animaux, occupant jusqu'à 60% des grands génomes de céréales (blé, orge ou maïs). Il est intéressant de constater que quelques familles de TEs peuvent représenter une grande partie de la multitude de TEs trouvés dans un génome.

Mécanisme de transposition des rétrotransposons

Les rétrotransposons sont les éléments les plus répandus dans les génomes du blé (prés de 60% de la séquence). Les sous-familles *Wis* et *Angela* de la famille *BARE-1 (Copia)* sont les plus fréquentes et représentent à elles seules de 10 à 20% du génome du blé (Charles et al., 2008). Les familles Sabrina (de la superfamille *Athila*) et *Fatima* (de la superfamille *Gypsy*) occupent chacune près de 7% du génome.

Les rétrotransposons autonomes codent une polyprotéine comprenant deux domaines : GAG et Pol. Ces deux domaines sont transcrits en une seule fois (Figure Annexe 4A, étape 1). La

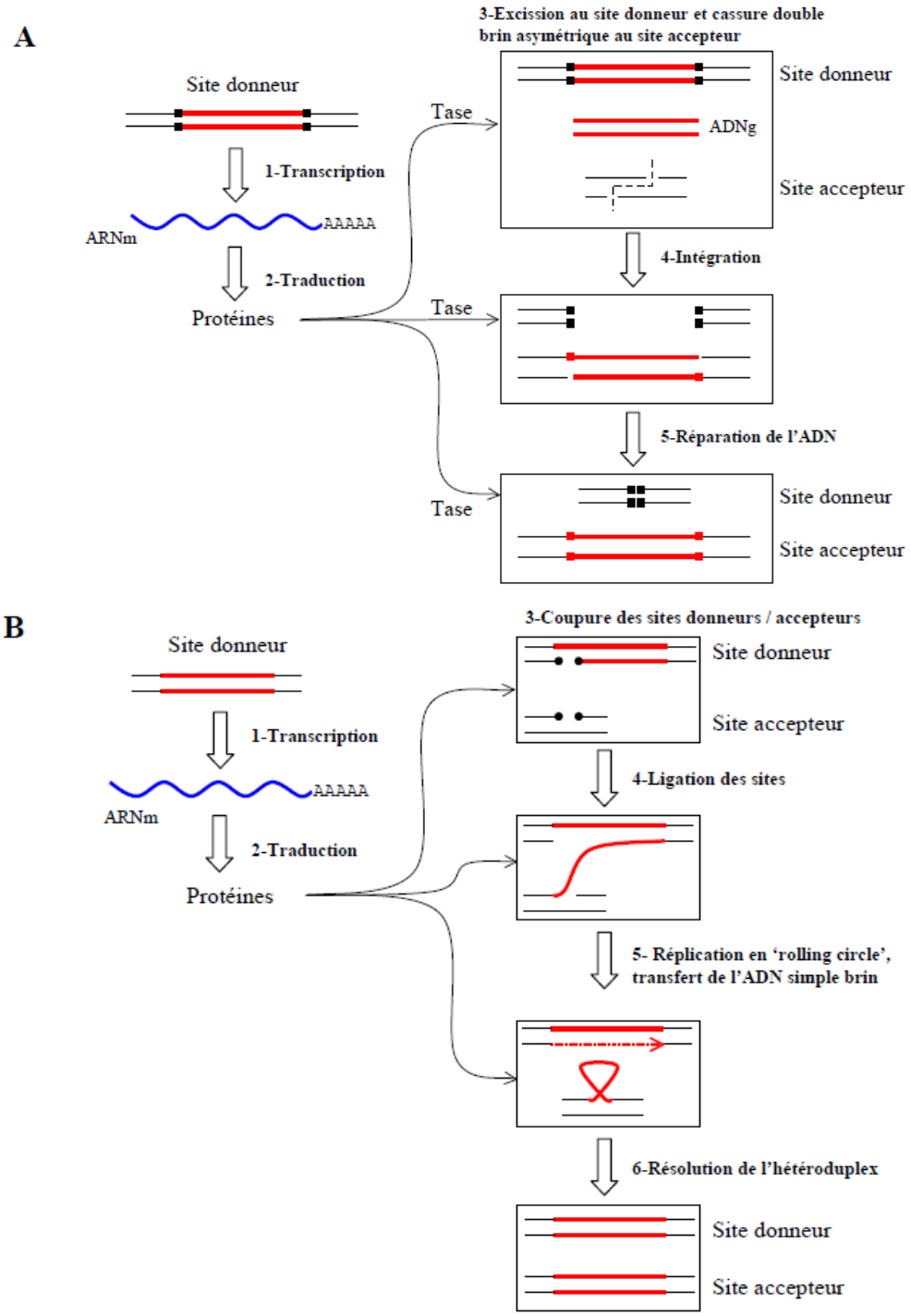


Figure Annexe 5: Mécanismes de transposition des principaux TEs de classe II, d'après (Sabot et al., 2004 ; Kapitonov and Jurka, 2006).

(A) Transposition des transposons à ADN de type *CACTA*. Les carrés noirs et rouges indiquent les TIRs. (B) Transposition des *hélitrons*. Les points noirs indiquent le site de cleavage entre A et T. Les différentes étapes des transpositions sont détaillées dans le texte correspondant.

traduction du messager donne la protéine GAG (protéine du capside) et le complexe de protéines Pol divisé en 4 protéines distinctes au niveau post-traductionnel (RT : transcriptase inverse, INT : intégrase, RH : RnaseH, AP : protéase aspartique) (Figure Annexe 4A, étape 2). Les protéines GAG vont se polymériser dans le cytoplasme pour former des VLP (Virus Like Particules) (Figure Annexe 4A, étape 3). Une partie de l'ARNm entre dans ces VLP, se dimérise et, sous l'action de la RH et la RT, est rétro-transcrit en ADN double brin (Figure Annexe 4A, étape 4). L'intégrase (INT) forme une cassure double brin de l'ADN génomique et y intégre l'ADN du rétrotransposon (Figure Annexe 4A, étape 5 et 6). La réparation de cette cassure asymétrique va former les TSD (Target Site Duplication), qui sont des signatures caractéristiques laissées par l'insertion de la plupart des TEs.

Mécanisme de transposition des rétroposons

Les rétroposons sont peu fréquents dans les génomes du blé (autour de 2%) et sont essentiellement de la famille des *LINE*.

Les rétroposons autonomes (tel que les *LINE*) ont un mécanisme de transposition assez différent de celui des rétrotransposons, bien que le principe général de la transposition soit le même : une transcription suivie d'une traduction des parties codantes de l'élément par la machinerie de l'hôte (Figure 5B, étape 1 et 2), puis la transcription reverse de l'ARN en ADNc, et la cassure asymétrique double brin de l'ADN au site cible suivi par l'intégration de l'élément. Une des différences principales, entre les deux mécanismes, est la transcription inverse qui a lieu sur le site d'insertion pour les rétroposons, et non dans une VLP.

Un rétroposon complet a généralement deux cadres ouverts de lecture (ORF pour Open Reading Frames) : ORF1 et ORF2. Le premier code pour une protéine pouvant se lier à l'ADN et le deuxième code pour une endonucléase (APE ou EN), une transcriptase inverse (RT) et parfois une RnaseH (RH).

L'endonucléase va permettre de lier un ARNm de l'élément avec la partie 3' d'un fragment d'ADN libre après une cassure double brin asymétrique, provoquée ou non par l'endonucléase (Figure 5B, étapes 3 et 4). La transcription inverse (action de RT) commence alors sur place, dans le sens 3' vers 5' (Figure Annexe 5B, étape 4). Elle se déroule cependant rarement en entier (10% des cas), produisant donc de nombreuses copies partielles de l'élément d'origine. Une fois la transcription terminée, l'ARNm est digéré (action de la RH) et la cassure double brin est réparée, dupliquant ainsi le rétroposon sur le deuxième brin d'ADN (Figure Annexe 5B, étape 5).

Mécanisme de transposition des transposons à ADN

Les transposons à ADN sont généralement moins invasifs que les rétrotransposons dans les génomes de plantes, car ils transposent par un système de 'couper-coller'. Cependant, ils constituent dans le blé jusqu'à 16% du génome dont la majorité appartient à la superfamille des *CACTA*. Une telle amplification serait expliquée par leur transposition à des moments précis du cycle cellulaire permettant d'augmenter leur nombre.

Les transposons à ADN complet de type *CACTA* ont deux ORF, le premier codant pour une transcriptase (Tase) et le second pour une protéine pouvant se lier à l'ADN mais dont le rôle reste à préciser (Wicker et al., 2003), les deux protéines sont transcrites et traduites par la machinerie de l'hôte (Figure 5A, étape 1 et 2). La transcriptase reconnaît spécifiquement les TIR (Tandem Inverted Repeat) présents aux extrémités de ces TE et catalyse toutes les étapes de la transposition, de l'excision à l'intégration (Figure Annexe 5A, étape 3, 4 et 5). La transposition se fait donc par un mécanisme de 'couper-coller', conservatif au niveau du nombre de copies. Mais si la transposition se produit en phase S du cycle cellulaire, en aval de la fourche de réplication, la cassure double brin provoquée par le transposon se répare en utilisant la chromatide sœur comme modèle possédant encore l'élément. L'élément est donc copié à un autre endroit du génome tout en gardant une copie à sa position d'origine.

Mécanisme de transposition des *Hélitrons*

Les *Hélitron*s (Kapitonov and Jurka, 2006) utilisent un mécanisme singulier, dit de « rolling-circle » en référence à un mécanisme similaire observé chez les géminivirus pour leur transposition. Ces éléments, transposant sans intermédiaire ARN, par un mécanisme original, et par 'copier-coller', étaient difficiles à décrire selon les critères de la première classification. Ils ont ainsi contribué à la formation de la nouvelle classification (Wicker et al., 2007). Contrairement à la plupart des autres éléments de classe II, les *Hélitron*s n'ont pas de TIR, motifs reconnus spécifiquement par les transposases. Leur séquence commence par TC, se termine par CTRR (R étant A ou G) et contient une séquence palindromique, de 16-20 pb située une dizaine de pb en amont du CTRR. Les *Hélitron*s s'insèrent entre les bases A et T d'un site accepteur.

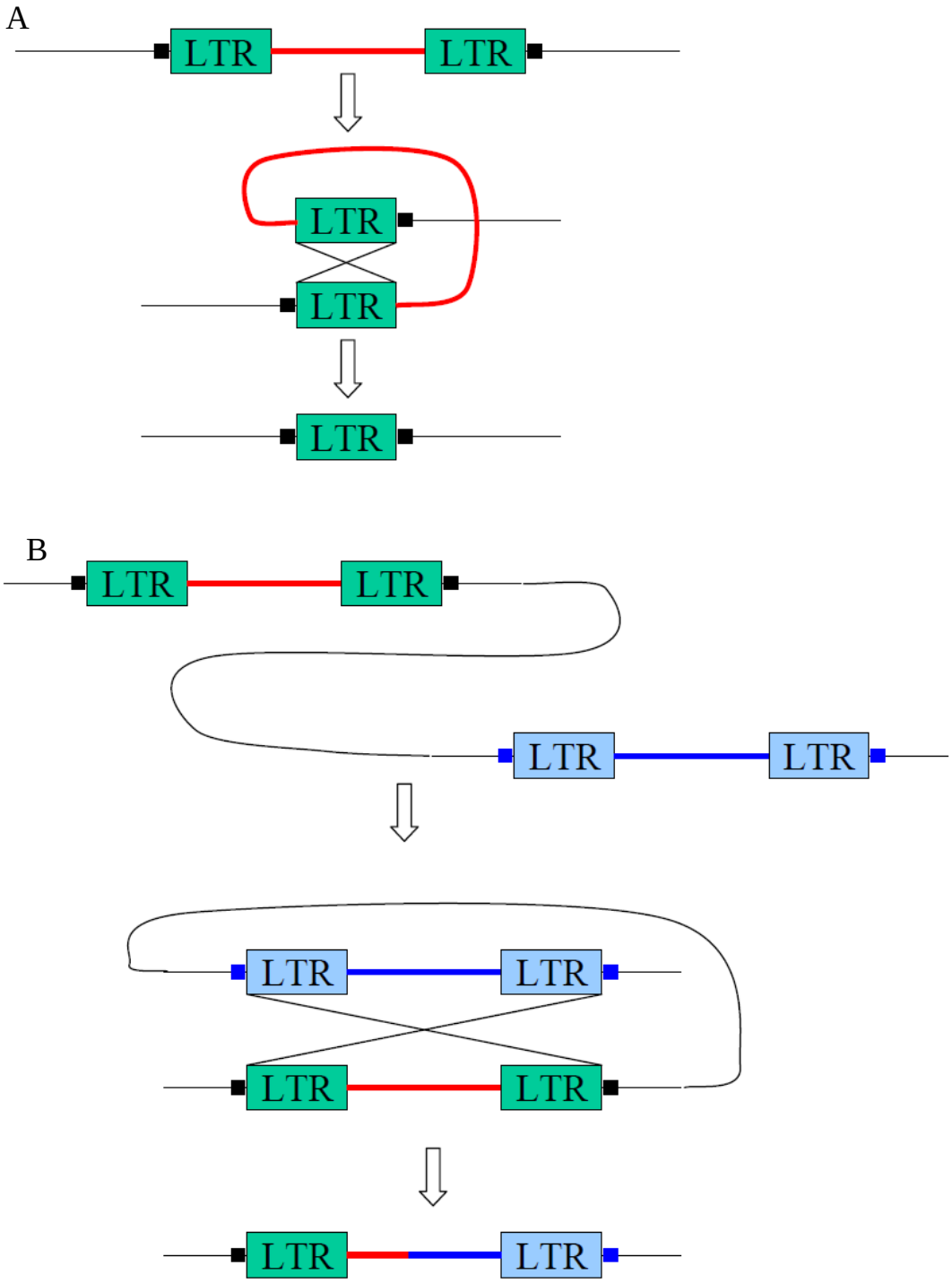


Figure Annexe 6: Mécanismes de délétion par recombinaison homologue inégale.

(A) Formation d'un Solo-LTR avec TSD par recombinaison homologue inégale entre les deux LTRs d'un rétrotransposon sur le même brin d'ADN. (B) Formation d'un rétrotransposon complet mais chimérique, sans TSD, par recombinaison entre deux éléments suffisamment similaires, sur le même brin d'ADN. Toute la séquence d'ADN entre les deux éléments est éliminée.

Selon les espèces, la composition des *Hélitron*s autonomes peut être assez variable, mais ils ont tous en commun un ORF codant pour la protéine 'RepHel' (Kapitonov and Jurka, 2006), constituée des domaines Rep (Réplicase) et Hel (Helicase).

Pour la transposition d'un *Hélitron*, suite à la transcription puis traduction des parties codantes de l'élément par la machinerie de l'hôte (Figure 5B, étape 1 et 2), le domaine Rep se lie aux sites donneur et accepteur, coupe l'ADN (simple brin) entre les bases A et T des deux sites (Figure 5B, étape 3) et fait une ligation entre la partie 3' du site donneur et la partie 5' du site accepteur (Figure 5B, étape 4). Le domaine Hel catalyse ensuite la synthèse de l'*Hélitron* par l'ADN polymérase de l'hôte au niveau du site donneur (Figure 5B, étape 5). En fin de synthèse, l'ADN simple brin correspondant à l'*Hélitron* original est transféré au site accepteur formant ainsi un hétéroduplex (Figure 5B, étape 5). Lors d'une prochaine réplication de l'ADN, cet hétéroduplex est résolu par la réplication de l'*Hélitron* sur le deuxième brin du site donneur (Figure 5B, étape 6). La séquence palindromique à la fin de l'*Hélitron* joue le rôle de terminateur de la réplication par 'rolling-circle'. Si cette séquence n'est pas bien reconnue, la synthèse continue sur le site donneur, si bien que le brin d'ADN transféré au site accepteur peut transporter non seulement l'*Hélitron*, mais aussi des séquences en 3' du site donneur. Ce phénomène a été notamment observé dans le maïs (Morgante et al., 2005) où des fragments de gènes voire des gènes entiers sont ainsi transportés à d'autres endroits du génome.

3.3. Mécanismes d'élimination des TEs

Les TEs ne restent pas intacts dans le génome hôte après leurs insertions, comme le montre le nombre important de copies tronquées et dégénérées trouvées dans les séquences génomiques disponibles. Les deux principaux mécanismes responsables de l'élimination des TEs sont les recombinaisons homologues inégales et les recombinaisons illégitimes (Vitte and Bennetzen., 2006; Devos, 2010).

Les recombinaisons homologues inégales font intervenir deux séquences suffisamment longues (généralement plusieurs centaines de pb) et similaires (>85-100% d'identité, selon la taille). Les rétrotransposons sont particulièrement concernés par ce type de recombinaison puisque leurs LTR correspondent à ces critères. La recombinaison peut ainsi avoir lieu entre les deux LTR d'un même élément (formation d'un Solo-LTR) (Figure 6A) ou les LTR d'éléments de la même famille (Figure 6B), aboutissant à la délétion plus ou moins importante de toute la séquence entre les deux éléments. Le taux de recombinaisons

homologues, révélées par la présence de Solo-LTR et d'éléments complets sans TSD, n'est pas très important dans les génomes du blé (taux de 1/50, (Charles et al., 2008)).
Les recombinaisons qui ne sont pas homologues sont dites 'illégitimes'. Le mécanisme de ces recombinaisons reste encore mal caractérisé. Elles impliquent souvent des motifs de quelques pb, conservés dans des orientations variables (directs, complémentaires, anti-sens). On observe ainsi des délétions allant de quelques pb à plusieurs dizaines de kb (Chantret et al., 2005). Menant à différents réarrangements génomiques, la recombinaison illégitime constitue un des mécanismes d'évolution majeurs chez le blé (Chantret et al., 2005 ; Charles et al., 2008).
Le locus Ha (hardness), chez le blé, contrôle la dureté du grain, la présence ou l'absence de ce locus explique la différence majeure qualitative entre le blé tendre (blé allohexaploïde) et le blé dur (allotétraploïde). Le locus Ha, représente trois gènes dupliqués en tandem (Pina et Pinb, Gsp1) qui contrôlent majoritairement le caractère «grain tendre» présents chez toutes les espèces diploïdes ancestrales du blé. Par des recombinaisons illégitimes, ce locus a été perdu dans les deux génomes A et B du blé tétraploïde (*T. turgidum*) menant au caractère « grain dur » (Chantret et al., 2005).

3.5. Le rôle des éléments transposables dans les génomes

Initialement identifiés comme des « éléments régulateurs » à leur découverte (McClintock, 1984), puis considérés comme des parasites du génome ou de l'ADN poubelle (Doolittle and Sapienza, 1980 ; Orgel and Crick, 1980) non nécessaires à l'organisme, les TEs, présents chez tous les organismes vivants, sont un des constituants les plus importants des génomes eucaryotes.

L'activité des TEs génère une diversité au sein d'une population et d'une espèce par une action directe ou indirecte. Les TEs affectent l'expression et la fonction des gènes, par leurs insertions dans des régions géniques (Lockton and Gaut, 2009), ou induisent des changements épigénétiques (méthylation des TE) (Parisod et al., 2010) et modifieraient la régulation de l'expression des gènes (Slotkin and Martienssen, 2007) et la stabilité du génome.

En effet, l'insertion d'un TE dans la séquence codante d'un gène va le plus souvent l'inactiver. Son insertion dans la zone promotrice du gène ou la présence d'un LTR, contenant des régions promotrices peut aussi altérer l'expression du gène (Kashkush et al.,

2003). La méthylation par l'organisme d'un TE peut également s'étendre jusqu'à un gène proche et conduire à une inactivation de la transcription de ce gène.

L'activation ou la répression de transposons spécifiques dans le polyploïde traduirait une adaptation du nouvel allopolyploïde. Les gènes fortement exprimés (up-régulés) dans l'allopolyploïde par rapport aux progéniteurs (transgressive upregulation) ou réprimés (mis sous silence) dans le polyploïde par rapport aux progéniteurs auraient un impact immédiat sur la physiologie du polyploïde et pourraient conduire à l'adaptation du nouveau polyploïde dans son environnement (Adams et al., 2003 ; Adams et al., 2004 ; Wang et al., 2004 ; Wang et al., 2006a ; Wang et al., 2006b ; Ni et al., 2009 ; Buggs et al., 2011a).

Les TEs peuvent également participer à la création de nouveaux exons ou même de nouveaux gènes. En effet, des études ont montré que certains éléments transposables (*Helitrons, PACK-Mule, CACTA* éléments de classe II) peuvent emporter des exons (ou même la totalité) des gènes adjacents lorsqu'ils transposent (Jiang et al., 2004 ; Lai et al., 2005). Ce phénomène « d'exon-shuffling », aboutit le plus souvent à des pseudogènes. Parfois, ces fragments sont capturés par un gène existant et peuvent former, dans de rares cas, un gène complètement nouveau (néo fonctionnalisation) (Jiang et al., 2004 ; Lai et al., 2005).

Lors de l'allopolyploidie, les séquences répétées spécifiques à l'une des espèces parentales subissent une élimination rapide (Han et al., 2005), ces séquences répétées d'ADN spécifiques représenteraient des fractions génomiques, principalement, incompatibles entre les deux génomes parentaux (Liu et al., 2009). Chez le tabac, cette élimination de séquences d'ADN répété spécifique dépendrait de mécanismes anciens conservés au cours de l'évolution (Skalicka et al., 2005).

Des études récentes ont montré l'implication des TEs au niveau de la forme et de la fonction des chromosomes par leurs insertions dans l'hétérochromatine autour des centromères et des télomères. Ils sont également impliqués dans la régulation de l'expression et les modifications de la chromatine par la voie des ARN interférents, RNAi (Slotkin and Martienssen, 2007).

Annexe 2: Pourcentage de lectures alignées avec le logiciel SOAP2 (Li et al, 2009).

Les pourcentages de lectures alignées présentées ont été obtenus sur le matériel de plantes étudié dans l'article présenté dans le chapitre 4.

Notation	Species	Ploidy level	Nb total of examined reads	Nb total of mapped reads	% of single aligned reads
D_t	Ae.tauschii 109 -11	diploid	39765951	4310867	10.84
D_t	Ae.tauschii 109 -14	diploid	74803165	8797587	11.76
D_t	Ae.tauschii 109 -16	diploid	83298821	7768899	9.32
B_jA_j	Joyau 11-2	tetraploid	24158972	2770592	11.47
B_jA_j	Joyau 12-2	tetraploid	21956365	2586967	11.78
B_jA_j	Joyau 15-1	tetraploid	19851102	2269758	11.32
B_hA_h	Tetra-Courtot BC6 -9	tetraploid	74342414	10648461	14.32
B_hA_h	Tetra-Courtot BC6 -13	tetraploid	68074206	9060109	13.30
B_hA_h	Tetra-Courtot BC6 -14	tetraploid	73274610	8092972	11.04
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	71398502	10463746	14.65
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	73576289	10046017	13.65
$B_hA_hD_t$	Synthetic TCx109 S1	hexaploid	82785477	8989309	10.85
$B_hA_hD_h$	Courtot	hexaploid	84639474	11575227	13.67
$B_hA_hD_h$	Courtot	hexaploid	81965434	9149574	11.16
$B_hA_hD_h$	Courtot	hexaploid	82617976	9258843	11.20

Références Bibliographiques

Adams, K., Percifield, R., and Wendel, J. (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics **168,** 2217-2226.

Adams, K., Cronn, R., Percifield, R., and Wendel, J. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ specific reciprocal silencing. Proc Natl Acad Sci USA **100,** 4649-4654.

Adams, K.L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. J Hered **98,** 136-141.

Adams, K.L., and Wendel, J. (2004). Exploring the genomic mysteries of polyploidy in cotton. Biological Journal of the Linnean Society **82,** 573-581.

Adams, K.L., and Wendel, J.F. (2005). Novel patterns of gene expression in polyploid plants. Trends Genet **21,** 539-543.

Adams, K.L., and Wendel, J.F. (2005b). Polyploidy and genome evolution in plants. Curr Opin Plant Biol **8,** 135-141.

Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L., Sasidharan, R., Reinke, V., Waterston, R., and Gerstein, M.B. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics **11,** 383.

Akhunov, E.D., Sehgal, S., Liang, H., Wang, S., Akhunova, A.R., Kaur, G., Li, W., Forrest, K.L., See, D., Simkova, H., Ma, Y., Hayden, M.J., Luo, M., Faris, J.D., Dolezel, J., and Gill, B.S. (2013). Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol **161,** 252-265.

Akhunova, A.R., Matniyazov, R.T., Liang, H., and Akhunov, E.D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics **11,** 505.

Al-Kaff, N., E. Knight, I. Bertin, T. Foote, N. Hart, S. Griffiths, and G. Moore. (2008). Detailed Dissection of the Chromosomal Region Containing the Ph1 Locus in Wheat Triticum aestivum: With Deletion Mutants and Expression Profiling. Annals of Botany **101,** 863-872.

Alkan, C., Kidd, J., Marques-Bonet, T., and al., e. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet **41,** 1061-1067.

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

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

Alwine, J.C., Kemp, D.J., and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A **74,** 5350-5354.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol **11,** R106.

Andreuzza, S., and Siddiqui, I. (2008). Spindle positioning, meiotic non reduction and polyploidy in plants. PLOS Genetics **4**.

APG , A.P.G. (1998). An ordinal classification for the families of flowering plants. . Annals of the Missouri Botanical Garden **85,** 531-553.

APG II, A.P.G. (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. Botanical Journal of the Linnean Society **141,** 399-436.

APG III, A.P.G. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society **161,** 105-121.

Arnaiz, O., Gout, J.F., Betermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E., and Sperling, L. (2010). Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics **11,** 547.

Arnaud, D., Chelaifa, H., Jahier, J., and Chalhoub, B. (2013). Reprogramming of Gene Expression in the Genetically Stable Bread Allohexaploid Wheat. In Polyploid and hybrid genomics., Z. Chen and J.A. Birchler, eds (Ames : Wiley-Blackwell), pp. 195-211.

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

Aubry, M. Principe des Tests Statistiques. Vocabulaire & Notions Générales (Rennes).

Axtell, M.J. (2013). Classification and comparison of small RNAs from plants. Annu Rev Plant Biol **64,** 137-159.

Bachem, C., van der Hoeven, RS., de Bruijn, SM., Vreugdenhil, D., Zabeau M, Visser RG. (1996). Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant J. **9,** 745-753.

Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., Baker, D., and al., e. (2011). Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nature Biotechnology **29,** 762 - 766.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y.Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet **56,** 406-414.

Barber, W.T., Zhang, W., Win, H., Varala, K.K., Dorweiler, J.E., Hudson, M.E., and Moose, S.P. (2012). Repeat associated small RNAs vary among parents and following hybridization in maize. **Proc Natl Acad Sci U S A. 109,** 10444-10449.

Bardil, A., Dantas de Almeida, J., Combes, M., Lashermes, P., and Bertrand, B. (2011). Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. New Phytologist **192,** 760-774.

Barroy-Hubler, F. (2003). Séquençage de génomes entiers. In Principes des techniques de biologie moléculaire., D. Tagu and C. Moussard, eds (Paris: Institut national de la recherche agronomique), pp. 176.

Baumel, A., Ainouche, M., Kalendar, R., and Schulman, A.H. (2002). Retrotransposons and genomic stability in populations of the young allopolyploid species Spartina anglica C.E. Hubbard (Poaceae). Mol Biol Evol **19,** 1218-1227.

Beaulieu, J., Jean, M., and Belzile, F. (2009). The allotetraploid Arabidopsis thaliana-Arabidopsis lyrata subsp. petraea as an alternative model system for the study of polyploidy in plants. Mol Genet Genomics **281,** 421-435.

Belcram, H. (2014). Organisation, évolution et fonctionnement des gènes majeurs de domestication (Q/q) chez les blés polyploïdes. In URGV-INRA (Evry: Evry University), pp. 270.

Bell, G.D., Kane, N.C., Rieseberg, L.H., and Adams, K.L. (2013). RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. Genome Biol Evol **5,** 1309-1323.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B **57,** 289-300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics **29,** 1165-1188.

Benjamini, Y., and Speed, T.P. (2011). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res **40,** e72.

Bennett, M.D., and Smith, J.B. (1976). Nuclear dna amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci **274,** 227-274.

Bennett, M.D., and Smith, J.B. (1991). Nuclear DNA amounts in angiosperms. . Philos.Trans. R. Soc. Lond. B. Biol. Sci. **334,** 309-345.

Bennetzen, J.L., and Kellogg, E.A. (1997). Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell **9,** 1509-1514.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Ann. Bot. (Lond.) **95,** 127-132.

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. Nat Rev Genet **14,** 333-346.

biostars. (2011). Burrow-Wheeler Alignment.

Birchler, J.A. (2012). Genetic Consequences of Polyploidy in Plants. In Polyploidy and Genome Evolution., P.S. Soltis and D.E. Soltis, eds (Berlin Heidelberg: Springer-Verlag), pp. 415.

Birchler, J.A., and Veitia, R.A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19,** 395-402.

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

Birchler, J.A., Auger, D.L., and Riddle, N.C. (2003). In search of the molecular basis of heterosis. Plant Cell **15,** 2236-2239.

Birchler, J.A., Bhadra, U., Bhadra, M.P., and Auger, D.L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol **234,** 275-288.

Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A. (2005). Dosage balance in gene regulation: biological implications. Trends Genet **21,** 219-226.

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

Blake, N.K., Lehfeldt, B.R., Lavin, M., and Talbert, L.E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome **42,** 351-360.

Blanc, G., and Wolfe, K.H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16,** 1679-1691.

Bloom, J., Khan, Z., Kruglyak, L., Singh, M., and Caudy, A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics **10,** 221.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. . Bioinformatics **19,** 185-193.

Bozdag, D., Barbacioru, C.C., and Catalyurek, U.V. (2009). Parallel Short Sequence Mapping for High Throughput Genome Sequencing. Int Parall Distrib P.**,** 1033-1042.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. (2003). Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A **100,** 3960-3964.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G., D'Amore, R., Allen, A., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., Kay, S., Waite, D., Trick, M., Bancroft, I., Gu, Y., Huo, N., Luo, M., Sehgal, S., Gill, B., Kianian, S., Anderson, O., Kersey, P., Dvorak , J., McCombie, W., Hall, A., Mayer, K., Edwards, K., Bevan, M., and Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. **491,** 705-710.

Bresson, A., Jorge V, Dowkiw A, Guerin V, Bourgait I, Tuskan GA, Schmutz J, Chalhoub B, Bastien C, and P., F.R. (2011). Qualitative and quantitative resistances to leaf rust finely mapped within two nucleotide-binding site leucine-rich repeat (NBS-LRR)-rich genomic regions of chromosome 19 in poplar. New Phytol **192,** 151-163.

Bretagnolle, F., and Thompson, J.D. (1995). Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. . TNew Phytologist **129,** 1-22.

Brownfield, L., and Kohler, C. (2011). Unreduced gamete formation in plants: mechanisms and prospects. J Exp Bot **62,** 1659-1668.

Buggs, R., Soltis, P., and Soltis, D. (2011a). Biosystematic relationships and the formation of polyploids. Taxon **60,** 324-332.

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

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

Buggs, R.J., Zhang, L., Miles, N., Tate, J.A., Gao, L., Wei, W., Schnable, P.S., Barbazuk, W.B., Soltis, P.S., and Soltis, D.E. (2011b). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol **21,** 551-556.

Burrows, M., and Wheeler, D. (1994). A block-sorting lossless data compression algorithm. In Digital Equipment Corporation, T. Rep., ed (Palo Alto (CA)).

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., and al., e. (2009). PASS: a program to align short sequences. Bioinformatics **25,** 967-968.

Capy, P. (1998). Classification of transposable elements. In Dynamics And Evolution Of Transposable Elements. , P. Capy, C. Bazin, D. Higuet, and T. Langin, eds (Austin: Landes Bioscience), pp. 37-52.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G.L., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., da Silva, M.L., Bockelman, H., Talbert, L., Anderson, J.A., Dreisigacker, S., Baenziger, S., Carter, A., Korzun, V., Morrell, P.L., Dubcovsky, J., Morell, M.K., Sorrells, M.E., Hayden, M.J., and

Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci U S A **110,** 8057-8062.

Chague, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., and Chalhoub, B. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol **187,** 1181-1194.

Chalabi, S., Chelaifa, H., Arnaud D, LeFloch E, Mestiri, I., DinhThi, V., LeClainche, I., Belcram, H., Devauchelle, C., Rizzon, C., Deffains, D., Huteau, V., Coriton, O., Chiquet, J., Jahier, J., and Chalhoub, C. (2014). Unraveling gene expression changes when decreasing and re-increasing alloploidy in wheat. in prepartion.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans B, Corréa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VH, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CH, Wang X, Canaguier A, Chauveau A, Bérard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Wang X, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, and Wincker, P. (2014). Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. Science **345,** 950-953

Chalupska, D., Lee, H.Y., Faris, J.D., Evrard, A., Chalhoub, B., Haselkorn, R., and Gornicki, P. (2008). Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci U S A **105,** 9691-9696.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P., and Chalhoub, B. (2005). Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). Plant Cell **17,** 1033-1045.

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

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

Charles, M. (2010). Évolution des génomes du blé (genres Aegilops et Triticum) au sein des Poaceae.

Dynamique rapide de l'espace occupé par les éléments transposables et conservation relative des gènes. In URGV-INRA (Evry: Evry University), pp. 314.

Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., Segurens, B., Carter, M., Huteau, V., Coriton, O., Appels, R., Samain, S., and Chalhoub, B. (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics **180,** 1071-1086.

Chase, M.W., and Reveal, J.L. (2009). A phylogenetic classification of the land plants to accompany APG III. Botanical Journal of the Linnean Society **161,** 122-127.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M., and Wendel, J.F. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). Genetics **182,** 503-517.

Chelaifa, H., Mahe, F., and Ainouche, M. (2010a). Transcriptome divergence between the hexaploid salt-marsh sister species Spartina maritima and Spartina alterniflora (Poaceae). Mol Ecol **19,** 2050-2063.

Chelaifa, H., Monnier, A., and Ainouche, M. (2010b). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina x townsendii and Spartina anglica (Poaceae). New Phytol **186,** 161-174.

Chelaifa, H., Chague, V., Chalabi, S., Mestiri, I., Arnaud, D., Deffains, D., Lu, Y., Belcram, H., Huteau, V., Chiquet, J., Coriton, O., Just, J., Jahier, J., and Chalhoub, B. (2013). Prevalence of gene expression additivity in genetically stable wheat allohexaploids. New Phytol **197,** 730-736.

Chen, F.Q., and Hayes, P.M. (1991). Wide hybridization of Hordeum vulgare x Zea mays. Genome **34,** 603-605.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., and al., e. (2009a). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods **6,** 677-U676

Chen, X. (2009). Small RNAs and their roles in plant development. Annu. Rev. Cell Dev. Biol. **25,** 21-44.

Chen, Y., Souaiaia, T., and Chen, T. (2009b). PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics **25,** 2514-2521.

Chen, Z.J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol **58,** 377-406.

Chen, Z.J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci **15,** 57-71.

Chen, Z.J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. Nat Rev Genet. **14,** 471-482.

Chen, Z.J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. Bioessays **28,** 240-252.

Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of Brassica plants. Nature **1**.

Chester, M., Gallagher, J.P., Symonds, V.V., Cruz da Silva, A.V., Mavrodiev, E.V., Leitch, A.R., Soltis, P.S., and Soltis, D.E. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). Proc Natl Acad Sci U S A **109,** 1176-1181.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., Leroy, P., Mangenot, S., Guilhot, N., Le Gouis, J., Balfourier, F., Alaux, M., Jamilloux, V., Poulain, J., Durand, C., Bellec, A., Gaspin, C., Safar, J., Dolezel, J., Rogers, J., Vandepoele, K., Aury, J.M., Mayer, K., Berges, H., Quesneville, H., Wincker, P., and Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. Science **345,** 1249721.

Clement, N.L., Snell, Q., Clement, M.J., Hollenhorst, P.C., Purwar, J., Graves, B.J., and al., e. (2010). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. . Bioinformatics **26,** 38-45.

Coate, J.E., Doyel, J.J. (2013). Genomics and Transcriptomics of Photosynthesis in Polyploids. In Polyploid and Hybrid Genomics, J. Chen and J.A. Birchler, eds (Wiley-Blackwell), pp. 384.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature **452,** 215-219.

Cole, C.G., McCann, O.T., Collins, J.E., Oliver, K., Willey, D., Gribble, S.M., Yang, F., McLaren, K., Rogers, J., Ning, Z., Beare, D.M., and Dunham, I. (2008). Finishing the finished human chromosome 22 sequence. Genome Biol **9,** R78.

Comai, L. (2000). Genetic and epigenetic interactions in allopolyploid plants. Plant Mol Biol **43,** 387-399.

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

Combes, M., Cenci, A., Baraille, H., Bertrand, B., and Lashermes, P. (2012). Homeologous gene expression in response to growing temperature in a recent allopolyploid (Coffea arabica L.). J Hered **103,** 36-46.

Combes, M., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid Coffea arabica grown at contrasted temperatures. New Phytol **200,** 251-260.

Consortium., T.T.G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485,** 635-641.

Coyne J.A, and H.A, O. (2004). Speciation. Sinauer. (Sunderland, MA).

Cui, C., Ge, X., Zhou, Y., Li, M., and Li, Z. (2013). Cytoplasmic and genomic effects on non-meiosis-driven genetic changes in Brassica hybrids and allotetraploids from pairwise crosses of three cultivated diploids. PLoS One **8,** e65078.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res **16,** 738-749.

Cusack, B., and Wolfe, K. (2007). When gene marriages don't work out: divorce by subfunctionalization. Trends Genet. **23,** 270-272.

d'Erfurth, I., Jolivet, S., Froger, N., Catrice, O., Novatchkova, M., and Mercier, R. (2009). Turning meiosis into mitosis. PLoS Biology **7,** e1000124.

d'Erfurth, I., Jolivet, S., Froger, N., Catrice, O., Novatchkova, M., Simon, M., Jenczewski, E., and Mercier, R. (2008). Mutations in AtPS1 (Arabidopsis thaliana parallel spindle 1) lead to the production of diploid pollen grains. . PLoS Genet **4**.

d'Erfurth, I., Cromer, L., Jolivet, S., Girard, C., Horlow, C., Sun, Y., To, J., Berchowitz, L., Copenhaver, G., and Mercier, R. (2010). The CYCLIN-A CYCA1;2/TAM is required for the meiosis I to meiosis II transition and cooperates with OSD1 for the prophase to first meiotic division transition. PLoS Genetics **6,** e1000989.

Darlington, C.D. (1937). Recent advances in Cytology. (London: Churchill.).

de Felippes, F., Wang, J., and Weigel, D. (2012). MIGS: miRNA-induced gene silencing. Plant J. **70,** 541-547.

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr Opin Plant Biol **15,** 168-176.

Deal, R., and Henikoff, S. (2011). Histone variants and modifications in plant gene regulation. Curr Opin Plant Biol **14,** 116-122.

Delmar, P., Robin, S., and Daudin, J.J. (2005). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics **21,** 502-508.

Devos, K. (2010). Grass genome organization and evolution. Curr Opin Plant Biol. **13,** 139-145.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., and Jaffrezic, F. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform **14,** 671-683.

Dixon, J., Braun, H., and Crouch, J. (2009). Transitioning wheat research to serve the future needs of the developing world. In Wheat facts and future., J. Dixon, H. Braun, and P. Kosina, eds (Mexico: D.F. CIMMYT), pp. 1-19.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res **36,** e105.

Dong, S., and Adams, K. (2011). Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. New Phytol **190,** 1045-1057.

Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature **284,** 601-603.

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

Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. Science **316,** 1862-1866.

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

Dvorak, J., Terlizzi, P., Zhang, H.B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome **36,** 21-31.

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

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

Eamens, A., and Waterhouse, P. (2011). Vectors and methods for hairpin RNA and artificial microRNAmediated gene silencing in plants. Methods Mol. Biol. **701,** 179-197.

Eaves, H.L., and Gao, Y. (2009). MOM: maximum oligonucleotide mapping. Bioinformatics **25,** 969-970.

Eckardt, N.A. (2000). Sequencing the rice genome. Plant Cell **12,** 2011-2017.

Erilova, A., Brownfield, L., Exner, V., Rosa, M., Twell, D., Mittelsten Scheid, O., Hennig, L., and Köhler, C. (2009). Imprinting of the Polycomb proup gene MEDEA serves as a ploidy sensor in Arabidopsis. PLoS Genetics **5,** e1000663.

Fang, Z., Martin, J., and Wang, Z. (2012). Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. Cell Biosci **2,** 26.

Farce, M. (2000). Génétique moléculaire, Productions animales, ISSN 0990-0632. (Institut National de la Recherche Agronomique).

Feldman, M., and Levy, A.A. (2009). Genome evolution in allopolyploid wheat--a revolutionary reprogramming followed by gradual changes. J Genet Genomics **36,** 511-518.

Feldman, M., Lupton, F.G.H., and Miller, T.E. (1995). Wheats. In Evolution of crops, J. Smartt and N.W. Simmonds, eds (London: Longman Scientific.

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

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A., and Vega, J.M. (1997). Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. Genetics **147,** 1381-1387.

Ferragina, P., and Manzin, i.G. (2000). Opportunistic data structures with applications. In 41st Annual Symposium on Foundations of Computer Science (Washington, DC), pp. 390-398.

Ferragina, P., and Manzini, G. (2005). Indexing compressed text. J ACM **52,** 552-581.

Feuillet, C., Le Gouis, J., and Charmet, G. (2011). Découvrir le génome du blé : un défi pour l'avenir. In Biotechnologie. (Clermont-Ferrand: UMR INRA-UBP Génétique, Diversité et Écophysiologie des Céréales.), pp. Biotechnologie.

Finch, R.A. (1983). Tissue-specific elimination of alternative whole parental genomes in one barley hybrid. Chromosoma **88,** 386-393.

Finnegan, D.J. (1990). Transposable elements and DNA transposition in eukaryotes. Curr Opin Cell Biol **2,** 471-477.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391,** 806-811.

Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. BMC Biol **6,** 16.

Flagel, L.E., and Wendel, J.F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol **186,** 184-193.

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

Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res **19,** 521-532.

Gaeta, R., Yoo, S., Pires, J., Doerge, R., Chen, Z., and Osborn, T. (2009). Analysis of gene expression in resynthesized Brassica napus Allopolyploids using arabidopsis 70mer oligo microarrays. PLoS One. **4,** e4760.

Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., and Osborn, T.C. (2007). Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell **19,** 3403-3417.

Gao, L., E. M. McCarthy, E. W. Ganko, and McDonald., J.F. (2004). Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics **5,** 18.

Garnis, C., Buys, T.P., and Lam, W.L. (2004). Genetic alteration and gene expression modulation during cancer progression. Mol Cancer **3,** 9.

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

Gates, R. (1908). The chromosomes of Oenothera. Science **27,** 193-195.

Gehring, M., Bubb, K.L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science **324,** 1447-1451.

Gernand, D., Rutten, T., Varshney, A., Rubtsova, M., Prodanovic, S., Brüss, C., Kumlehn, J., Matzk, F., and Houben, A. (2005). Uniparental Chromosome Elimination at Mitosis and Interphase in Wheat and Pearl Millet Crosses Involves Micronucleus Formation, Progressive Heterochromatinization, and DNA Fragmentation. Plant Cell. **17,** 2431-2438.

Ghildiyal, M., and Zamore, P. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet **10,** 94-108.

Gianinetti, A. (2013). A criticism of the value of midparent in polyploidization. J Exp Bot **64,** 4119-4129.

Goldblatt, P. (1980). Polyploidy in angiosperms: Monocotyledons. In Polyploidy: Biological relevance, L.W. H., ed (New York, USA: Plenum Press New York), pp. 219-239.

Gong, L., Kakrana, A., Arikit, S., Meyers, B.C., and Wendel, J.F. (2013). Composition and expression of conserved microRNA genes in diploid cotton (Gossypium) species. Genome Biol Evol **5,** 2449-2459.

Grandbastien, M.A., Audeon, C., Bonnivard, E., Casacuberta, J.M., Chalhoub, B., Costa, A.P., Le, Q.H., Melayah, D., Petit, M., Poncet, C., Tam, S.M., Van Sluys, M.A., and Mhiri, C. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res **110,** 229-241.

Grant, V. (1981). Plant speciation. (Columbia University Press).

Green, E. (2001). Strategies for the systematic sequencing of complex genomes. Nature Reviews Genetics **2,** 573-583.

Griffiths, A., Gelbart, W., Miller, J., and Lewontin, R. (1999). Chromosomal Rearrangements. In Modern Genetic Analysis (New York: W. H. Freeman.

Griffiths, S., Sharp, R., Foote, T., Bertin, I., Wanous, M., Reader, S., Colas, I., and Moore, G. (2006). Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature **439,** 749-752.

Groszmann, M., Paicu, T., Alvarez , J., Swain, S., and Smyth, D. (2011). SPATULA and ALCATRAZ, are partially redundant, functionally diverging bHLH genes required for Arabidopsis gynoecium and fruit development. Plant J. **68,** 816-829.

Grover, C.E., Yu, Y., Wing, R.A., Paterson, A.H., and Wendel, J.F. (2008). A phylogenetic analysis of indel dynamics in the cotton genus. Mol Biol Evol **25,** 1415-1428.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homeolog expression bias and expression level dominance in allopolyploids. The New Phytologist. **196,** 966-971.

Gu, Z., Nicolae, D., Lu, H., and Li, W. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18,** 609-613.

Guan, X., Pang, M., Nah, G., Shi, X., Ye, W., Stelly, D.M., and Chen, Z.J. (2014). miR828 and miR858 regulate homoeologous MYB2 gene functions in Arabidopsis trichome and cotton fibre development. Nat Commun **5,** 3050.

Ha, M., Kim, E.D., and Chen, Z.J. (2009a). Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci U S A **106,** 2295-2300.

Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K., Chapman, E., Carrington, J., Chen, X., Wang, X., and Chen, Z. (2009b). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci U S A. **106,** 17835-17840.

Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E., and Sahinalp, S. (2010). mrsFast: a cache-oblivious algorithmfor short-read mapping. Nat Methods **7,** 576-577.

Han, F., Fedak, G., Guo, W., and Liu, B. (2005). Rapid and repeatable elimination of a parental genome-specific DNA repeat (pGc1R-1a) in newly synthesized wheat allopolyploids. Genetics **170,** 1239-1245.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res **38,** e131.

Harlan, J., and deWet, J. (1975). The origins of polyploidy. The Botanical Review **41,** 361-390

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. Science **320,** 106-109.

Haston, E., Richardson, J.E., Stevens, P.F., Chase, M.W., Harris, D.J. (2009). The Linear Angiosperm Phylogeny Group (LAPG) III : a linear sequence of the families in APG III. Botanical Journal of the Linnean Society **161,** 128-131.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A., and F., W.J. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res **16,** 1252-1261.

He, P., Friebe, B.R., Gill, B.S., and Zhou, J.M. (2003). Allopolyploidy alters gene expression in the highly stable hexaploid wheat. . Plant Mol Biol **52,** 401-414.

Hegarty, M., Barker, G., Wilson, I., Abbott, R., Edwards, K., and Hiscock, S. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Curr Biol. **16,** 1652-1659.

Hegarty, M.J., Jones, J.M., Wilson, I.D., Barker, G.L., Coghill, J.A., Sanchez-Baracaldo, P., Liu, G., Buggs, R.J., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2005). Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. Mol Ecol **14,** 2493-2510.

Henson, J., Tischler, G., and Ning, Z. (2012). Next-generation sequencing and large genome assemblies. Pharmacogenomics **13,** 901-915.

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Galvez, S., Schaaf, S., Jouve, N., Simkova, H., Valarik, M., Dolezel, J., and 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. Plant J **69,** 377-386.

Heslop-Harrison, J.S. (2012). Genome evolution: extinction, continuation or explosion? Curr Opin Plant Biol **15,** 115-121.

Higgins, J., Magusin, A., Trick, M., Fraser, F., and Bancroft, I. (2012). Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species Brassica napus. BMC Genomics. **13,** 247.

Hochberg, Y., and Tamhane, A. (1987). Multiple comparison procedures. (New-York: Wiley & Sons).

Homer, N., Merriman, B., and Nelson, S.F. (2009). BFAST: an alignment tool for large scale genome resequencing. PLoS One **4,** e7767.

Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform **11,** 181-197.

Hovav, R., Udall, J.A., Chaudhary, B., Rapp, R., Flagel, L., and Wendel, J.F. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc Natl Acad Sci U S A **105,** 6191-6195.

Hsieh, T.F., Ibarra, C.A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L., and Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. Science **324,** 1451-1454.

Hu, Z., Han, Z., Song, N., Chai, L., Yao, Y., Peng, H., Ni, Z., and Sun, Q. (2013). Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (Triticum aestivum). New Phytol **197,** 1344-1352.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002a). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proceedings of the National Academy of Sciences, USA **99,** 8133-8138.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002b). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc Natl Acad Sci U S A **99,** 8133-8138.

Huet, S. (2013). «Il faut ouvrir la boîte noire du génome des blés». In Liberation Science.

Hufton, A.L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. Curr Opin Genet Dev **19,** 600-606.

IHGSC. (2004). Finishing the euchromatic sequence of the human genome. Nature **431,** 931-945.

Ilut, D.C., Coate, J.E., Luciano, A.K., Owens, T.G., May, G.D., Farmer, A., and Doyle, J.J. (2012). A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am J Bot **99,** 383-396.

Ingram, R., and Noltie, H. (1995). Senecio cambrensis Rosser, Biological Flora of the British Isles. . J Ecol **83,** 537-546.

Initiative, T.A.G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408,** 796-815.

Initiative., I.B. (2010). Genome sequencing and analysis of themodel grass Brachypodium distachyon. Nature **463,** 763-768

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and P., S.T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4,** 249-264.

IWGSC. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science **345,** 1251788.

Jackson, S., and Chen, Z.J. (2010). Genomic and expression plasticity of polyploidy. Curr Opin Plant Biol **13,** 153-159.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., and Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449,** 463-467.

Jeanmougin, M. (2012). Statistical methods for robust analysis of transcriptome data by integration of biological knowledge. . In Statistique & Génome (Evry: Val d'Essonne), pp. 172.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS One. **5,** e12336.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K.F., Li, D., Pan, S., Zheng, F., Hu, Q., Xia, X., Li, J., Liang, Q., Chen, J., Wicker, T., Gou, C., Kuang, H., He, G., Luo, Y., Keller, B., Xia, Q., Lu, P., Wang, J., Zou, H., Zhang, R., Xu, J., Gao, J., Middleton, C., Quan, Z., Liu, G., Wang, J., Yang, H., Liu, X., He, Z., Mao, L., and Wang, J. (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature **496,** 91-95.

Jiang, H., and Wong, W.H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics **24,** 2395-2396.

Jiang, N., Bao, Z., Zhang, X., Eddy, S., and Wessler, S. (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature **431,** 569-573.

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., and dePamphilis, C.W. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473,** 97-100.

Jones, D.C., Ruzzo, W.L., Peng, X., and Katze, M.G. (2012). A new approach to bias correction in RNA-Seq. Bioinformatics **28,** 921-928.

Juliano, C., Wang, J., and Lin, H. (2011). Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annu. Rev. Genet. **45,** 447-469.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110,** 462-467.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., Ashburner, M., and Celniker, S.E. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol **3,** RESEARCH0084.

Kapitonov, V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. Proc Natl. Acad. Sci. **103,** 4540-4545.

Kasha, K.J., and Kao, K.N. . (1970). High frequency haploid production in barley (Hordeum vulgare L.). Nature **225,** 874-875.

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

Kashkush, K., M. Feldman, and A. A. Levy. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics **160,** 1651-1659.

Katiyar-Agarwal, S., Gao, S., Vivian-Smith, A., and Jin, H. (2007). A novel class of bacteria-induced small RNAs in Arabidopsis. Genes Dev **21,** 3123-3134.

Keane, O.M., Toft, C., Carretero-Paulet, L., Jones, G.W., and Fares, M.A. (2014). Preservation of genetic and regulatory robustness in ancient gene duplicates of Saccharomyces cerevisiae. Genome Res.

Kellogg, E.A. (2001). Evolutionary history of the grasses. Plant Physiol **125,** 1198-1205.

Kenan-Eichler, M., Leshkowitz D, Tal L, Noor E, Melamed-Bessudo C, and al, e. (2012). Wheat hybridization and polyploidization results in deregulation of small RNAs. Genetics **188,** 263-272.

Kerber, E.R. (1964). Wheat: Reconstitution of the Tetraploid Component (AABB) of Hexaploids. Science **143,** 253-255.

Khasdan, V., B. Yaakov, Z. Kraitshtein, and K. Kashkush. . (2010). Developmental timing of DNA elimination following allopolyploidization in wheat. Genetics **185,** 387-390.

Kidwell, M.G. (2002). Transposable elements and the evolution of genome size in eukaryotes. Genetica **115,** 49-63.

Kihara, H. (1944). Discovery of the DD-analyser, one of the ancestors of vulgare wheats. Ag. Hort. (Tokyo) **19,** 889-890.

Kim, Y.J., Teletia, N., Ruotti, V., Maher, C.A., Chinnaiyan, A.M., Stewart, R., and al., e. (2009). ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches. Bioinformatics. **25,** 1424-1425.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. Nat Methods **3,** 211-222.

Koh, J., Soltis, P.S., and Soltis, D.E. (2010). Homeolog loss and expression changes in natural populations of the recently and repeatedly formed allotetraploid Tragopogon mirus (Asteraceae). BMC Genomics **11,** 97.

Książczyk, T., Kovarik, A., Eber, F., Huteau, V., Khaitova, L., and al, e. (2011). Immediate unidirectional epigenetic reprogramming of NORs occurs independently of rDNA rearrangements in synthetic and natural forms of a polyploid species Brassica napus. Chromosoma **120,** 557-571

Lackey, E., Ng, D.W., and Chen, Z.J. (2010). RNAi-mediated down-regulation of DCL1 and AGO1 induces developmental changes in resynthesized Arabidopsis allotetraploids. New Phytol **186,** 207-215.

Lai, J., Li, Y., Messing, J., and and Dooner, H.K. (2005). Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proceedings of the National Academy of Sciences of the United States of America **102,** 9068-9073.

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

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J.,

Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y.J. (2001). Initial sequencing and analysis of the human genome. Nature **409,** 860-921.

Landry, C.R., Wittkopp, P.J., Taubes, C.H., Ranz, J.M., Clark, A.G., and Hartl, D.L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics **171,** 1813-1822.

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods **9,** 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10,** R25.

Laurie, D.A., and Bennett, M.D. (1988). Cytological evidence for fertilization in hexaploid wheat x sorghum crosses. Plant Breed. **100,** 73-82.

Laurie, D.A., and Bennett, M.D. . (1986). Wheat x maize hybridization. Can. J. Genet. Cytol. **28,** 313-316.

Lavania, U., Srivastava, S., Lavania, S., Basu, S., Misra, N., and Mukai, Y. (2012). Autopolyploidy differentially influences body size in plants, but facilitates enhanced accumulation of secondary metabolites, causing increased cytosine methylation. Plant J **71,** 539-549.

Lecrom, S., and Marc, P. (2011). DNA microarray principle.

Lee, H.S., and Chen, Z.J. (2001). Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. Proc Natl Acad Sci U S A **98,** 6753-6758.

Lee, W.P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., and Marth, G.T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One **9,** e90581.

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

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature **396,** 643-649.

Levy, A., and Feldman, M. (2004). Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. Biological Journal of the Linnean Society **82,** 607-613.

Levy, A.A., and Walbot, V. (1990). Regulation of the timing of transposable element excision during maize development. Science **248,** 1534-1537.

Levy, A.A., and Feldman, M. (2002). The impact of polyploidy on grass genome evolution. Plant Physiol **130,** 1587-1593.

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

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25,** 1754-1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26,** 589-595.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18,** 1851-1858.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25,** 1966-1967.

Li, W., Huang L, and BS, G. (2008b). Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. Plant Physiol. **146,** 200-212.

Liang, P., and Pardee, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science **257,** 967-971.

Lim, K.Y., Soltis, D.E., Soltis, P.S., Tate, J., Matyasek, R., Srubarova, H., Kovarik, A., Pires, J.C., Xiong, Z., and Leitch, A.R. (2008). Rapid chromosome evolution in recently formed polyploids in Tragopogon (Asteraceae). PLoS One **3,** e3353.

Lin, H., Zhang, Z., Zhang, M.Q., Ma, B., and Li, M. (2008). ZOOM! Zillions of oligos mapped. . Bioinformatics **24,** 2431-2437.

Lindner, R., and Friedel, C.C. (2012). A comprehensive evaluation of alignment algorithms in the context of RNA-seq. PLoS One **7,** e52403.

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., Gao, C., Wu, H., Li, Y., Cui, Y., Guo, X., Zheng, S., Wang, B., Yu, K., Liang, Q., Yang, W., Lou, X., Chen, J., Feng, M., Jian, J., Zhang, X., Luo, G., Jiang, Y., Liu, J., Wang, Z., Sha, Y., Zhang, B., Wu, H., Tang, D., Shen, Q., Xue, P., Zou, S., Wang, X., Liu, X., Wang, F., Yang, Y., An, X., Dong, Z., Zhang, K., Zhang, X., Luo, M.C., Dvorak, J., Tong, Y., Wang, J., Yang, H., Li, Z., Wang, D., Zhang, A., and Wang, J. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature **496,** 87-90.

Liu, B., and Wendel, J.F. (2003). Epigenetic phenomena and the evolution of plant allopolyploids. Mol Phylogenet Evol **29,** 365-379.

Liu, B., and Davis, T.M. (2011). Conservation and loss of ribosomal RNA gene sites in diploid and polyploid Fragaria (Rosaceae). BMC Plant Biol **11,** 157.

Liu, B., J. M. Vega, and Feldman., M. (1998). Rapid genomic changes in newly synthesized amphiploids of Triticum and Aegilops. II. Changes in low-copy coding DNA sequences. Genome **41,** 535-542.

Liu, B., Brubaker, C.L., Mergeai, G., Cronn, R.C., and Wendel, J.F. (2001). Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome **44,** 321-330.

Liu, B., Xu, C., Zhao, N., Qi, B., Kimatu, J.N., Pang, J., and Han, F. (2009). Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. J Genet Genomics **36,** 519-528.

Liu, B., J. M. Vega, G. Segal, S. Abbo, M. Rodova, and M. Feldman. . (1998). Rapid genomic changes in newly synthesized amphiploids of Triticum and Aegilops. I. Changes in low copy non-coding DNA sequences. Genome **41,** 272-277.

Liu, C., Wong, T., Wu, E., Luo, R., Yiu, S., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., Li, R., and Lam, T. (2012a). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics **28,** 878-879.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012b). Comparison of next-generation sequencing systems. J Biomed Biotechnol **2012,** 251364.

Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2010). A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res **39,** 578-588.

Lockton, S., and Gaut, B.S. (2009). The contribution of transposable elements to expressed coding sequence in Arabidopsis thaliana. J Mol Evol **68,** 80-89.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C., Krieger, M., Scott, M., Zipursky, S., and Darnell, J. (2005). Biologie moléculaire de la cellule. (Bruxelles).

Lu, Y., Arnaud, D., Belcram, H., Falentin, C., Rouault, P., Piel N, Lucas MO, Just J, Renard M, Delourme R, and B., C. (2012). A Dominant Point Mutation in a RINGv E3 Ubiquitin Ligase Homoeologous Gene Leads to Cleistogamy in Brassica napusC. Plant Cell. **24,** 4875-4879.

Lukens, L.N., Pires, J.C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napus allopolyploids. Plant Physiol **140,** 336-348.

Luo, M.C., Gu, Y.Q., You, F.M., Deal, K.R., Ma, Y., Hu, Y., Huo, N., Wang, Y., Wang, J., Chen, S., Jorgensen, C.M., Zhang, Y., McGuire, P.E., Pasternak, S., Stein, J.C., Ware, D., Kramer, M., McCombie, W.R., Kianian, S.F., Martis, M.M., Mayer, K.F., Sehgal, S.K., Li, W., Gill, B.S., Bevan, M.W., Simkova, H., Dolezel, J., Weining, S., Lazo, G.R., Anderson, O.D., and Dvorak, J. (2013). A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. Proc Natl Acad Sci U S A **110,** 7940-7945.

Lutz, A. (1907). A preliminary note on the chromosomes of Oenothera Lamarckiana and one of its mutants, O. gigas. Science **26,** 151-152.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics **154,** 459-473.

Ma, J., Stiller, J., Berkman, P.J., Wei, Y., Rogers, J., Feuillet, C., Dolezel, J., Mayer, K.F., Eversole, K., Zheng, Y.L., and Liu, C. (2013). Sequence-based analysis of translocations and inversions in bread wheat (Triticum aestivum L.). PLoS One **8,** e79329.

Mach, J. (2011). Whole-Genome Duplications: Does Metabolic Connectivity Influence Gene Retention? The Plant Cell **23,** 1683

Madlung, A., and Wendel, J.F. (2013). Genetic and epigenetic aspects of polyploid evolution in plants. Cytogenet Genome Res **140,** 270-285.

Madlung, A., Masuelli, R.W., Watson, B., Reynolds, S.H., Davison, J., and Comai, L. (2002). Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic Arabidopsis allotetraploids. Plant Physiol **129,** 733-746.

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

Malinska, H., Tate, J.A., Matyasek, R., Leitch, A.R., Soltis, D.E., Soltis, P.S., and Kovarik, A. (2010). Similar patterns of rDNA evolution in synthetic and recently formed natural populations of Tragopogon (Asteraceae) allotetraploids. BMC Evol Biol **10,** 291.

Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol **9,** 34.

Marcussen, T., Sandve, S., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, Jakobsen, K., Wulff, B., Steuernagel, B., Mayer,

K., and Olsen, O. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science **345,** 1250092.

Marguerat, S., Wilhelm, B.T., and Bahler, J. (2008). Next-generation sequencing: applications beyond genomes. Biochem Soc Trans **36,** 1091-1096.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437,** 376-380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18,** 1509-1517.

Marone, D., Laido, G., Gadaleta, A., Colasuonno, P., Ficco, D.B., Giancaspro, A., Giove, S., Panio, G., Russo, M.A., De Vita, P., Cattivelli, L., Papa, R., Blanco, A., and Mastrangelo, A.M. (2012). A high-density consensus map of A and B wheat genomes. Theor Appl Genet **125,** 1619-1638.

Martin-Magniette, M.L. (2012). Analyse des données d'expression issues de la puce CATMAv6: normalisation, analyse différentielle. (INRA-URGV), pp. 10.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat Rev Genet **12,** 671-682.

Matkovich, S.J., Zhang, Y., Van Booven, D.J., and Dorn, G.W., 2nd. (2010). Deep mRNA sequencing for in vivo functional analysis of cardiac transcriptional regulators: application to Galphaq. Circ Res **106,** 1459-1467.

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

Matsushita, S., Tyagi, A., Thornton, G., Pires, J., and Madlung, A. (2012). Allopolyploidization lays the foundation for evolution of distinct populations: evidence from analysis of synthetic Arabidopsis allohexaploids. Genetics **191,** 535-547.

Matzk, F. (1996). Hybrids of crosses between oat and Andropogene or Paniceae species. . Crop Sci. **36,** 17-21.

Matzk, F., and Mahn, A. (1994). Improved techniques for haploid production in wheat using chromosome elimination. Plant Breed **113,** 125-129.

Matzk, F., Oertel, C., Altenhofer, P., and Schubert, I. (1997). Manipulation of reproductive systems in Poaceae to increase the efficiency in crop breeding and production. Trends Agron. **1,** 19-34.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci U S A **74,** 560-564.

Mayer, K.F., and al., e. (2012). A physical, genetic, and functional sequence assembly of the barley genome. Nature **491,** 711-716.

Mayer, K.F.X., and al., e. (2009). Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. **151,** 496-505.

Mayfield, D., Chen, Z.J., and Pires, J.C. (2011). Epigenetic regulation of flowering time in polyploids. Curr Opin Plant Biol **14,** 174-178.

McClintock, B. (1984). The significance of responses of the genome to challenge. . Science **226,** 792-801.

McFadden, E.S., and Sears, E.R. (1946). The origin of Triticum speltoides and its free-threshing hexaploid relatives. J. Hered. **37,** 81-89.

Mestiri, I., Chague, V., Tanguy, A.M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Chalhoub, B., and Jahier, J. (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. New Phytol **186,** 86-101.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R.E., Michael, T.P., Wall, K., Rice, D.W., Albert, H., Wang, M.L., Zhu, Y.J., Schatz, M., Nagarajan, N., Acob, R.A., Guan, P., Blas, A., Wai, C.M., Ackerman, C.M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.K., Shakirov, E.V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J.E., Gschwend, A.R., Delcher, A.L., Singh, R., Suzuki, J.Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M.J., Feltus, F.A., Porter, B., Li, Y., Burroughs, A.M., Luo, M.C., Liu, L., Christopher, D.A., Mount, S.M., Moore, P.H., Sugimura, T., Jiang, J., Schuler, M.A., Friedman, V., Mitchell-Olds, T., Shippen, D.E., dePamphilis, C.W., Palmer, J.D., Freeling, M., Paterson, A.H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature **452,** 991-996.

Misra, S., Narayanan, R., Lin, S., and al., e. (2009). FANGS: high speed sequence mapping for next generation sequencers. In ACM Symposium on Applied Computing (Sierre, Switzerland), pp. 1539-1546.

Morgante, M., S. Brunner, G. Pea, K. Fengler, and A. Zuccolo, a.A.R. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies
diversity in maize. . Nat. Genet. **37,** 997-1002.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods **5,** 621-628.

Muntzing, A. (1936). The evolutionary significance of autopolyploidy. . Hereditas **21,** 263-378.

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

Nesbitt, M., and Samuel, D. (1996). From the staple crop to extinction? The archaeology and history of hulled wheats. . . In Hulled Wheats. Proceedings of the First International Workshop on Hulled Wheats. (International Plant Genetic Resources Institute, Rome, Italy.).

Ni, Z., Kim ED, Ha M, Lackey E, and Liu J, e.a. (2009). Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature **457,** 327-331.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Res. **11,** 1725-1729.

Nobuta, K., Lu, C., Shrivastava, R., Pillay, M., DePaoli, E., and al., a. (2008). Distinct size distributionof endogeneous siRNAs in maize: evidence from deep sequencing in the mop1-1 mutant. Proc. Natl. Acad. Sci. USA **105,** 14958-14963.

Nomura, T., Ishihara, A., Yanagita, R.C., Endo, T.R., and Iwamura, H. (2005). Three genomes differentially contribute to the biosynthesis of benzoxazinones in hexaploid wheat. Proc Natl Acad Sci U S A **102,** 16490-16495.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature **388,** 167-171.

Nyren, P., Pettersson, B., and Uhlen, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem **208,** 171-175.

Orgel, L.E., and Crick, H.C. (1980). SelfishDNA: The ultimate parasite. Nature **284,** 604-607.

Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., and Martienssen, R.A. (2003). Understanding mechanisms of novel gene expression in polyploids. Trends Genet **19,** 141-147.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct **4,** 14.

Ossowski, S., Schwab, R., and Weigel, D. (2008). Gene silencing in plants using artificial microRNAs and other small RNAs. Plant J. **53,** 674-690.

Otto, S.P. (2007). The evolutionary consequences of polyploidy. Cell **131,** 452-462.

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

Ozkan, H., Tuna, M., and Galbraith, D. (2006). No DNA loss in autotetraploids of Arabidopsis thaliana. Plant Breed **125,** 288-291.

Page, J., Huynh, M., Liechty, Z., Grupp, K., Stelly, D., Hulse, A., Ashrafi, H., Van Deynze, A., Wendel, J., and Udall, J. (2013a). Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing. G3 (Bethesda). **3,** 1809-1818.

Page, J.T., Gingle, A.R., and Udall, J.A. (2013b). PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 (Bethesda) **3,** 517-525.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science **315,** 241-244.

Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M., and Chen, Z.J. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome Biol **10,** R122.

Pareek, C.S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. J Appl Genet **52,** 413-435.

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M., and Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol **184,** 1003-1015.

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

Parry, E.M., and Cox, B.S. (1970). The tolerance of aneuploidy in yeast. Genet Res **16,** 333-340.

Paterson, A.H., and al., e. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457,** 551-556.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A **101,** 9903-9908.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.J., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.H., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., ur Rahman, M., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.K., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A., Zhang, T., Dennis, E.S., Mayer, K.F., Peterson, D.G., Rokhsar, D.S., Wang, X., and Schmutz, J. (2011). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature **492,** 423-427.

Paux, E., Roger, D., Badaeva, E., Gay, G., Bernard, M., Sourdille, P., and Feuillet, C. (2006). Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J **48,** 463-474.

Pearson, B., Gaskin, D., Segers, R., Wells, J., Nuijten, P., and van Vliet, A. (2007). The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828). J Bacteriol **189,** 8402-8403.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. (2000). Evidence for DNA loss as a determinant of genome size. Science **287,** 1060-1062.

Pfeifer, M., Kugler, K.G., Sandve, S.R., Zhan, B., Rudi, H., Hvidsten, T.R., Mayer, K.F., and Olsen, O.A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. Science **345,** 1250091.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, H. Kim, K. Collura, D. S. Brar, S. Jackson, R. A. Wing, and Panaud., O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. . Genome Res **16,** 1262-1269.

Pihan, G., and Doxsey, S.J. (2003). Mutations and aneuploidy: co-conspirators in cancer? Cancer Cell **4,** 89-94.

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

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Dolezel, J., Fahima, T., Budak, H., Keller, B., Salvi, S., Maccaferri, M.,

Steinbach, D., Feuillet, C., Quesneville, H., and Salse, J. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J **76,** 1030-1044.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet **3,** 827-837.

Pritham, E.J. (2009). Transposable elements and factors influencing their success in eukaryotes. J. Hered. **100,** 648-655.

Project., A.G. (2013). The Amborella genome and the evolution of flowering plants. Science **342,** 1241089.

Project., I.R.G.S. (2005). The map-based sequence of the rice genome. Nature **436,** 793–800

Pumphrey, M., Bai, J., Laudencia-Chingcuanco, D., Anderson, O., and Gill, B.S. (2009). Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics **181,** 1147-1157.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F., and Liu, B. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. BMC Biol **10,** 3.

Quail, M., Kozarewa, I., Smith, F., Scally, A., Stephens, P., Durbin, R., Swerdlow, H., and Turner, D. (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods **5,** 1005-1010.

Ragupathy, R., and Cloutier, S. (2008). Genome organisation and retrotransposon driven molecular evolution of the endosperm Hardness (Ha) locus in Triticum aestivum cv Glenlea. Mol Genet Genomics **280,** 467-481.

Rambani, A., Page, J.T., and Udall, J.A. (2014). Polyploidy and the petal transcriptome of Gossypium. BMC Plant Biol **14,** 3.

Ramsey, J., and Schemske, D.W. (1998). Pathways, Mechanisms, an Rates of polyploid formation in flowering plants. Annual Review Of Ecology And Systematics **29,** 467-501.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol **14,** R95.

Rapp, R.A., and Wendel, J.F. (2005). Epigenetics and plant evolution. New Phytol **168,** 81-91.

Rapp, R.A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. BMC Biol **7,** 18.

Raskina, O., A. Belyayev, and E. Nevo. . (2002). Repetitive DNas of wild emmer wheat (Triticum dicoccoides) and their relation to S-genome species: molecular cytogenetic analysis. Genome **45,** 391-401.

Ren, J., Sun, D., Chen, L., You, F.M., Wang, J., Peng, Y., Nevo, E., Sun, D., Luo, M.C., and Peng, J. (2013). Genetic diversity revealed by single nucleotide polymorphism markers in a worldwide germplasm collection of durum wheat. Int J Mol Sci **14,** 7061-7088.

Riddle, N.C., and Birchler, J.A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. Trends Genet **19,** 597-600.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol Evol **16,** 351-358.

Riley, R., and Chapman, V. (1958). Genetic control of the cytologically diploid behaviour of hexaploid wheat. Nature **182,** 713-715.

Riley, R., Unrau, J., and al., e. (1958). Evidence on the origin of the B genome of wheat. J. Hered. **49,** 91-98.

Rines, H.W., and Dahleen, L.S. . (1990). Haploids of plants produced by application of maize pollen to emasculated oat florets. Crop Sci. **30,** 1073-1078.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. BMC Bioinformatics **12,** 480.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol **12,** R22.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. Science **281,** 363, 365.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem **242,** 84-89.

Rostoks, N., Mudie, S., Cardle, L., Russell, J., Ramsay, L., Booth, A., Svensson, J.T., Wanamaker, S.I., Walia, H., Rodriguez, E.M., Hedley, P.E., Liu, H., Morris, J., Close, T.J., Marshall, D.F., and Waugh, R. (2005). Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics **274,** 515-527.

Rothberg, J., Hinz, W., Rearick, T., Schultz, J., Mileski, W., Davey, M., Leamon , J., Johnson, K., Milgrew, M., Edwards, M., Hoon, J., Simons, J., Marran, D., Myers, J., Davidson, J., Branting, A., Nobile, J., Puc, B., Light, D., Clark , T., Huber, M., Branciforte, J., Stoner, I., Cawley, S., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J., Namsaraev, E., McKernan, K., Williams, A., Roth, G., and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature **475,** 348-352.

Roulin, A., Auer, P.L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R.W., and Jackson, S.A. (2013). The fate of duplicated genes in a polyploid plant genome. Plant J.

Ruiz-Ferrer, V., and Voinnet, O. (2009). Roles of plant small RNAs in biotic stress responses. Annu. Rev. Plant Biol. **60,** 485-510.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. **5,** e1000386.

Ryther, R.C., et al. (2004). Gh1 splicing is regulated by multiple enhancers whose mutation produces a dominantnegative

gh isoform that can be degraded by allele-specific small interfering rna (sirna). Endocrinology **145,** 2988-2996.

Sabot, F.o., Simon, D., and Bernard, M. (2004). Plant transposable elements, with an emphasis on grass species. . Euphytica **139,** 227-247.

Saha, S., Sparks, AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. (2002). Using the transcriptome to annotate the genome. Nat Biotechnol. **20,** 508-512.

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

Salmon, A., Ainouche, M.L., and Wendel, J.F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol **14,** 1163-1175.

Salse, J., S. Bolot, M. Throude, V. Jouffe, B. Piegu, U. Masood Quraishi, T. Calcagno, R. Cooke, M. Delseny, and C. Feuillet. (2008a). Identification and Characterization of

Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. Plant Cell. **20**
11-24.

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

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74,** 5463-5467.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and Bennetzen., J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. **20,** 43-45.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., MelakeBerhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science **274,** 765-768.

Saze, H., Shiraishi, A., Miura, A., and Kakutani, T. (2008). Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. Science **319,** 462-465.

Schatz, M.C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics **25,** 1363-1369.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.F. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J Comput Biol **19,** 796-813.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C. (2004). Mining EST databases to resolve evolutionary events in major crop species. Genome **47,** 868-876.

Schnable, J., and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias towards one of two maize subgenomes. PLoS One **6,** e17855.

Schnable, J., Springer, N., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. **108,** 4069-4074.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S.,

Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326,** 1112-1115.

Schreiber, A.W., Hayden, M.J., Forrest, K.L., Kong, S.L., Langridge, P., and Baumann, U. (2012). Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat. BMC Genomics **13,** 492.

Semon, M., and Wolfe, K.H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. Proc Natl Acad Sci U S A **105,** 8333-8338.

Service, R.F. (2006). Gene sequencing - The race for the $1000 Genome. Science **311,** 1544-1546.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and 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.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int **2014,** 309650.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science **309,** 1728-1732.

Shi, X., Ng, D.W., Zhang, C., Comai, L., Ye, W., and Chen, Z.J. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. 3**,** 950.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A **89,** 8794-8797.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet **15,** 121-132.

Skalicka, K., Lim, K.Y., Matyasek, R., Matzke, M., Leitch, A.R., and Kovarik, A. (2005). Preferential elimination of repeated DNA sequences from the paternal, Nicotiana tomentosiformis genome donor of a synthetic, allotetraploid tobacco. New Phytol **166,** 291-303.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet **8,** 272-285.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzic, M., Becker, J.D., Feijo, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell **136,** 461-472.

Slotte, T., Huang HR, Holm K, Ceplitis A, Onge KS, Chen J, Lagercrantz U, and M., L. (2009). Splicing variation at a FLOWERING LOCUS C homeolog is associated with flowering time variation in the tetraploid Capsella bursa-pastoris. Genetics **183,** 337-345.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., and al., e. (2009). Updates to the RMAP short-read mapping software. Bioinformatics **25,** 2841-2842.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments.

. Statistical Applications in Genetics and Molecular Biology **3**.

Soltis, D.E., and al., e. (2013). Polyploidy and Genome evolution. In The early stages of polyploidy: rapid and repeated evolution in Tragopogon, P.S. Soltis and D.E. Soltis, eds (Heidelberg: Springer.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2008). Polyploidy and angiosperm diversification. Am J Bot **96,** 336-348.

Soltis, P.S., and Soltis, D.E. (2000). The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci U S A **97,** 7051-7057.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu Rev Plant Biol **60,** 561-588.

Soltis, P.S., and Soltis, D.E. (2012). Polyploidy and Genome Evolution. (Berlin Heidelberg: Springer-Verlag).

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics **14,** 91.

Song, K., Lu, P., Tang, K., and Osborn, T.C. (1995). Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc Natl Acad Sci U S A **92,** 7719-7723.

Southern, E. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology **98,** 503-517.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Res **6,** 2601-2610.

Stamati, K., Mackay, I., and Powell, W. (2009). A quantitative genomic imbalance gene expression assay in a hexaploid species: wheat (Triticum aestivum). Genome **52,** 89-94.

Stebbins, G.L. (1950). Variation and evolution in plants. (New York: Columbia University Press).

Stekel, D. (2003). Microarray Bioinformatics. (New York: Cambridge University Press).

Stupar, R., Bhaskar, P., Yandell, B., Rensink, W., and Hart, A., et al. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. Genetics **176,** 2055-2067.

Stupar, R.M., and Springer, N.M. (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. Genetics **173,** 2199-2210.

Su, Y.A., Wu, J., Zhang, L., Zhang, Q., Su, D.M., He, P., Wang, B.D., Li, H., Webster, M.J., Rennert, O.M., and Ursano, R.J. (2008). Dysregulated mitochondrial genes and networks with drug targets in postmortem brain of patients with posttraumatic stress disorder (PTSD) revealed by human mitochondria-focused cDNA microarrays. Int J Biol Sci **4,** 223-235.

Subrahmanyam, N.C. (1977). Haploidy from Hordeum interspecific crosses. I. Polyhaploids of H. parodii and H. procerum. Theor. Appl. Genet. **49,** 209-217.

Tagu, D., and Moussard, C. (2003). Principes des techniques de biologie moléculaire. (INRA, Editions Quae).

Takumi, S., Nasuda, S., and al., e. (1993). Wheat phylogeny determined by RFLP analysis of nuclear DNA. I. Einkorn wheat. Jpn. J. Genet. **68,** 73-79.

Talbert, L.E., Blake, N.K., Storlie, E.W., and Lavin, M. (1995). Variability in wheat based on low-copy DNA sequence comparisons. Genome **38,** 951-957.

Tanaka, T., Kobayashi, F., Joshi, G.P., Onuki, R., Sakai, H., Kanamori, H., Wu, J., Simkova, H., Nasuda, S., Endo, T.R., Hayakawa, K., Dolezel, J., Ogihara, Y., Itoh, T., Matsumoto, T., and Handa, H. (2013). Next-Generation Survey Sequencing and the Molecular Organization of Wheat Chromosome 6B. DNA Res **21,** 103-114.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science **320,** 486-488.

Tang, H., Woodhouse, M., Cheng, F., Schnable, J., Pedersen, B., Conant, G., Wang, X., Freeling, M., and Pires, J. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics **190,** 1563-1574.

Tang, H., X. Wang, J. E. Bowers, R. Ming, M. Alam, and A. H. Paterson. . (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18,** 1944-1954.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. Genome Res **21,** 2213-2223.

Tate, J.A., Joshi, P., Soltis, K.A., Soltis, P.S., and Soltis, D.E. (2009). On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid Tragopogon miscellus (Asteraceae). BMC Plant Biol **9,** 80.

Tate, J.A., Ni, Z., Scheen, A.C., Koh, J., Gilbert, C.A., Lefkowitz, D., Chen, Z.J., Soltis, P.S., and Soltis, D.E. (2006). Evolution and expression of homeologous loci in Tragopogon miscellus (Asteraceae), a recent and reciprocally formed allopolyploid. Genetics **173,** 1599-1611.

Tawfik, D.S., and Griffiths, A.D. (1998). Man-made cell-like compartments for molecular evolution. Nat Biotechnol **16,** 652-656.

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

Team, R.D.C. (2008). R: A language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing).

Tian, L., Li, X., Ha, M., Zhang, C., and Chen, Z. (2014). Genetic and epigenetic changes in a genomic region containing MIR172 in Arabidopsis allopolyploids and their progenitors. Heredity (Edinb) **112,** 207-214.

Torres, E., Williams, B., and Amon, A. (2008). Aneuploidy: Cells Losing Their Balance. Genetics. **179,** 737-746.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. . Bioinformatics **25,** 1105-1111.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-

Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechn **28,** 511-515.

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

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313,** 1596-1604.

Udall, J.A., Quijada, P.A., and Osborn, T.C. (2005). Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of Brassica napus L. Genetics **169,** 967-979.

Udall, J.A., Swanson, J.M., Nettleton, D., Percifield, R.J., and Wendel, J.F. (2006). A novel approach for characterizing expression levels of genes duplicated by polyploidy. Genetics **173,** 1823-1827.

USDA, f., RFI, BRGP, INRA. (2012). Production française de blé.

Van de Peer, Y., Maere, S., and Meyer, A. (2009a). The evolutionary significance of ancient genome duplications. Nat Rev Genet **10,** 725-732.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009b). The flowering world: a tale of duplications. Trends Plant Sci **14,** 680-688.

Vedel, F., F. Quetier, Y. Cauderon, F. Drosba and G. Doussinault. . (1981). Studies on maternal inheritance in polyploid wheat with cytoplasmic DNAs as genetic markers. . Theoretical and Applied Genetics **59,** 239-245.

Veitia, R.A. (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics **168,** 569-574.

Veitia, R.A., Bottani, S., and Birchler, J.A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet **24,** 390-397.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M.,

Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Dematte, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J.A., Sterck, L., Vandepoele, K., Grando, S.M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S.K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F., and Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One **2,** e1326.

Verhoeven, K., Jansen, J., van Dijk, P., and Biere, A. (2010). Stress-induced DNA methylation changes and their heritability in asexual dandelions. New Phytologist **185,** 1108-1118.

Vitte, C., and Bennetzen., a.J.L. (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad.Sci. USA **103,** 17638-17643.

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. Clin Chem **55,** 641-658.

Vogel, e.a. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature **463,** 763-768.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell **136,** 669-687.

Wang, J., Tian, L., Lee, H.S., and Chen, Z.J. (2006a). Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in Arabidopsis allopolyploids. Genetics **173,** 965-974.

Wang, J., Tian L, Madlung A, Lee HS, Chen M, and al., e. (2004). Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. . Genetics **167,** 1961-1973.

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

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., and Chen, Z.J. (2006b). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172,** 507-517.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K., Wang, B., Song, C., Shi, N., Kohel, R.J., Percy, R.G., Yu, J.Z., Zhu, Y.X., Wang, J., and Yu, S. (2012). The draft genome of a diploid cotton Gossypium raimondii. Nat Genet **44,** 1098-1103.

Waterhouse, P.M., Wang, M.-B., and Lough, T. (2001). Gene silencing as an adaptive defence against viruses. Nature **411,** 834-842.

Wei, F., Zhang, J., Zhou, S., He, R., Schaeffer, M., Collura, K., Kudrna, D., Faga, B.P., Wissotski, M., Golser, W., Rock, S.M., Graves, T.A., Fulton, R.S., Coe, E., Schnable, P.S., Schwartz, D.C., Ware, D., Clifton, S.W., Wilson, R.K., and Wing, R.A. (2009). The physical and genetic framework of the maize B73 genome. PLoS Genet **5,** e1000715.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol **42,** 225-249.

Wendel, J.F., and Doyle, J.J. (2005). Polyploidy and Evolution in plants. In Plant Diversity and Evolution, R.J. Henry, ed (Wallingford, UK: CABI Publishing), pp. 97-117.

Wendel, J.F., Cronn, R.C., Johnston, J.S., and Price, H.J. (2002). Feast and famine in plant genomes. Genetica **115,** 37-47.

Whitley, P., Lemire, A., Brockman, J., Sheila Heater, S., Schageman, J., Gu, J., Lea, K., Qu, L., San Jose, C., Hernandez, N., Bramlett, K., Ilsley, D., and Setterquist, R. (2009). A Comparison of Next Generation Sequencing and Microarrays for Whole Transcriptome Expression Profiling (Life Technologies/Ambion R&D, 2130 Woodward, Austin, TX, USA, 78744).

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z. D. Liu, J. Dubcovsky, and Keller., B. (2003). Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. . Plant Cell **15,** 1186-1197.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A.H. (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet **8,** 973-982.

Wicker, T., Krattinger, S., Lagudah, E., Komatsuda, T., Pourkheirandish, M., Matsumoto, T., Cloutier S, Reiser L, Kanamori H, Sato K, Perovic D, Stein N, and B, K. (2009). Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. Plant Physiol. **149,** 258-270.

Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L., Coghill, J.A., Burridge, A., Hall, A., Brenchley, R.C., D'Amore, R., Hall, N., Bevan, M.W., Richmond, T., Gerhardt, D.J., Jeddeloh, J.A., and Edwards, K.J. (2012). Targeted re-sequencing of the allohexaploid wheat exome. Plant Biotechnol J **10,** 733-742.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2,** 333-341.

Wright, K.M., Pires, J.C., and Madlung, A. (2009). Mitotic instability in resynthesized and natural polyploids of the genus Arabidopsis (Brassicaceae). Am J Bot **96,** 1656-1664.

Wu, T., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **7,** 873-881.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association **99,** 909-917.

Xiong, Z., Gaeta, R.T., and Pires, J.C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci U S A **108,** 7908-7913.

Xu, C., Bai, Y., Lin, X., Zhao, N., Hu, L., Gong, Z., Wendel, J.F., and Liu, B. (2014). Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. Mol Biol Evol **31,** 1066-1076.

Xu, Y., Zhong L, Wu X, Fang X, and J, W. (2009). Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta **229,** 471-483.

Yaakov, B., and Kashkush, K. (2011). Methylation, transcription, and rearrangements of transposable elements in synthetic allopolyploids. Int J Plant Genomics **2011,** 569826.

Yang, T., Furuta, Y., Nagata, S., and Watanabe, N. (1999). Tetra Chinese Spring with AABB genomes extracted from the hexaploid common wheat, Chinese Spring. . Genes & Genetic Systems **74,** 67-70.

Yoo, M.J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) **110,** 171-180.

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

Zenkteler, M., and Nitzsche, W. (1984). Wide hybridization experiments in cereals. Theor. Appl. Genet. **68,** 311-315.

Zhang, H., Bian, Y., Gou, X., Zhu, B., Xu, C., Qi, B., Li, N., Rustgi, S., Zhou, H., Han, F., Jiang, J., von Wettstein, D., and Liu, B. (2013). Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. Proc Natl Acad Sci U S A **110,** 3447-3452.

Zhang, H., Zhu, B., Qi, B., Gou, X., Dong, Y., Xu, C., Zhang, B., Huang, W., Liu, C., Wang, X., Yang, C., Zhou, H., Kashkush, K., Feldman, M., Wendel, J.F., and Liu, B. (2014). Evolution of the BBAA Component of Bread Wheat during Its History at the Allohexaploid Level. Plant Cell.

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., and Chalhoub, B. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci U S A **108,** 18737-18742.

Zhao, N., Zhu, B., Li, M., Wang, L., Xu, L., Zhang, H., Zheng, S., Qi, B., Han, F., and Liu, B. (2011). Extensive and heritable epigenetic remodeling and genetic stability accompany allohexaploidization of wheat. Genetics **188,** 499-510.

Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One **9,** e78644.

Zheng, W., Chung, L.M., and Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. BMC Bioinformatics **12,** 290.

Zhou, R., Moshgabadi, N., and Adams, K.L. (2011). Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. Proc Natl Acad Sci U S A **108,** 16122-16127.

Zuccolo, A., A. Sebastian, J. Talag, Y. Yu, H. Kim, K. Collura, D. Kudrna, and Wing., R.A. (2007). Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol. Biol. **7,** 152.

Résumé

La polyploïdie ou la duplication des génomes est une force majeure dans l'évolution et l'adaptation des espèces, notamment des angiospermes qui ont tous eu des évènements de polyploïdisation réccurrents au cours de leur évolution. Afin de comprendre la reprogrammation de l'expression des gènes en réponse à la polyploïdie chez les espèces économiquement importantes du blé (genres *Triticum* et *Aegilops*), j'ai utilisé un modèle original, qui consiste à caractériser les réponses à la diminution puis la ré-augmentation du niveau de ploïdie. Ainsi, le blé allotétraploïde (*T. turgidum*, BBAA) est extrait à partir du blé naturel allohexaploïde (*T. aestivum*, BBAADD). Ce blé allotétraploïde extrait est hybridé à son tour à l'espèce diploïde *Ae. tauschii* (DD) pour synthétiser un blé allohexaploïde.

J'ai utilisé des méthodes d'analyse de l'expression des gènes basées sur les microarrays ainsi que le séquençage massif des ARN (RNA-Seq), basé sur les outils de nouvelles générations de séquençage (NGS) et rendu possible par la récente mise à disposition des séquences de trois copies homéologues (A_h, B_h, D_h) de 8605 gènes. Les méthodes bioinformatiques et statistiques appropriées ont été développées et/ou utilisées.

Mes travaux révèlent un partitionnement de l'expression des gènes en celles des homéologues qui les composent dans les différents allopolyploïdes étudiés. La majorité des homéologues contribuent à l'expression globale des gènes de manière équivalente (1/3 chacun), d'autres présentent un biais d'expression spécifique vers un des homéologues, sans montrer de dominance d'un des sous-génomes. Une concertation dans le partitionnement et le niveau d'expression des homéologues est bien établie dans le blé: la majorité des homéologues augmentent leur expression lorsqu'ils sont séparés et la diminuent lorsqu'ils sont rassemblés dans un niveau de ploïdie supérieur. Ainsi, dans le blé allohexaploïde, pour la majorité des gènes, l'expression des homéologues A_h et B_h est égale au 2/3 de leur niveau d'expression dans le blé allotétraploïde extrait, et l'expression de l'homéoallèle D_h est égale au 1/3 du niveau d'expression dans le blé diploïde donneur du génome D. Cette concertation de l'expression des homéologues maintiendrait l'expression globale des gènes à des niveaux similaires dans les différentes espèces de blé de différents niveaux de ploïdie.

Les résultats obtenus contribuent à la compréhension de la régulation de l'expression des gènes dans les polyploïdes du blé et la contribution des homéologues qui les composent. Les analyses futures des fonctions des différentes catégories d'expression des gènes permettraient d'identifier des fonctions particulières régulées en réponse à la polyploïdie.

Mots-clefs: blé, allopolyploïdie, homéologues, reprogrammation de l'expression, séquençage.

Abstract

Polyploidy is a major evolutionary force, especially in angiosperms, all of which species have undergone recurrent polyploïdization events during their evolution.

In order to understand reprogramming of gene expression in response to polyploidy in the economically important wheat species (genera *Triticum* and *Aegilops*), I used an original model that consists in decreasing and reincreasing ploidy levels. Thus, the allotetraploid *T. turgidum* (BBAA) is extracted from the allohexaploid bread wheat *T. aestivum* (BBAADD), consisting in decreasing ploidy level. This extracted allotetraploid is crossed with the diploid species *Ae. tauschii* (DD) to synthesize an allohexaploid wheat, consisting in re-increasing ploidy level.

The characterization of reprogramming of gene expression in response to decreasing and re-increasing ploidy levels was done here using first microarray technologies and then massive parallel mRNA sequencing (RNA-Seq), that has been rendered possible by the recent 'draft' hexaploid wheat genome sequencing and subsequently the availability of the three homoelogs sequences (A_h, B_h, D_h) of 8605 genes. Adequate bioinformatics and statistics methods have been adopted and/or developed and used.

My work reveals a partitioning of global expression of genes into that of their constituent homoeologs in different wheats allopolyploids. Most of homoeologs contribute equally to the overall gene expression and a low proportion reveals a bias towards one homoeolog, without showing a global dominance of a specific sub-genome. The partitioning and concerted expression of homoeologs is also established in wheat. Most homoeologs increase their expression when separated and reduce their expression levels when joined together in a higher ploidy level. For most genes, A_h and B_h homoeolog expression in allohexaploid wheat is equal to 2/3 of their expression level in the extracted allotetraploid wheat whereas the D_h homoeolog expression level is equal to 1/3 of that in the wheat diploid genome. This concerted change in homoeolog expression maintains the global gene expression at nearly similar levels in different ploidy levels.

Results obtained in this work contribute to our understanding of global gene expression regulation and its partitioning between constituent homoeologs at different ploidy levels. Functional analysis of the different gene expression categories would reveal important gene functional categories that are regulated in response to polyploidy.

Key-words: wheat, allopolyploidy, homoeologs, reprogramming of gene expression, sequencing.