Université Blaise Pascal de Clermont-Ferrand *Année 2010*

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

THESE

Présentée à l'Université Blaise Pascal
Pour l'obtention du titre de Docteur d'Université
Spécialité : Physiologie et génétique moléculaires

Camille Rustenholz

Impact de la structure du génome sur l'organisation, la régulation et la fonction des gènes sur le chromosome 3B du blé hexaploïde (*Triticum aestivum* L.).

Soutenue le 15 décembre 2010 devant la commission d'examen :

Rapporteurs : H. QUESNEVILLE, Directeur de Recherches INRA, Versailles-Grignon
O. PANAUD, Professeur à l'Université de Perpignan Via Domitia
T. HEULIN, Directeur de recherche CNRS, CEA, Cadarache
Examinateurs : S. MOUZEYAR, Professeur à l'Université Blaise Pascal, Clermont-Ferrand
N. YAHIAOUI, Chargée de Recherches CIRAD, Montpellier
Encadrants : C. FEUILLET, Directrice de Recherches INRA, Clermont-Ferrand
E. PAUX, Chargé de Recherches INRA, Clermont-Ferrand

UMR 1095 INRA-UBP « Génétique, Diversité et Ecophysiologie des Céréales »
234 Avenue de Brézet – 63100 Clermont-Ferrand

Résumé

Du fait de sa taille (17 Gb), de sa nature allohexaploïde et de son fort taux de séquences répétées (>80%), le génome du blé tendre a toujours été considéré comme trop complexe pour des analyses moléculaires efficaces. En conséquence, la connaissance de la structure de son génome reste limitée. Utilisant une approche chromosome-spécifique, la carte physique du chromosome 3B du blé a récemment été établie et a permis le développement de ressources génomiques uniques. Pendant ma thèse, la mise en œuvre d'approches transcriptomiques utilisant ces ressources m'a permis d'analyser les relations entre la structure du génome, l'évolution, la fonction et la régulation des gènes le long du chromosome 3B de blé.

Tout d'abord, des filtres portant les BAC du « Minimal Tiling Path » (MTP) du chromosome 3B ont été hybridés avec 15 échantillons d'ARNm pour identifier les BAC portant des gènes. Ensuite, des puces Agilent 15K d'expression d'orge ont été hybridées avec les pools tridimensionnels (3D) du MTP du chromosome 3B pour localiser les gènes plus précisément sur la carte physique. Afin de construire la première carte transcriptionnelle d'un chromosome de blé, ces mêmes pools 3D ainsi que les 15 échantillons d'ARNm ont été hybridés sur des puces NimbleGen 40K d'expression de blé. Les résultats obtenus à partir de ces expériences ont permis de tirer des conclusions quant à l'organisation de l'espace génique sur le chromosome 3B. Ainsi les gènes sont répartis tout le long du chromosome 3B selon un gradient de densité de gènes du centromère vers les télomères avec une plus forte proportion de gènes regroupés en îlots au niveau des télomères. Une analyse évolutive a montré que les îlots seraient essentiellement constitués de gènes ayant subi des réarrangements dans le génome du blé. De plus, la carte transcriptionnelle a également mis en évidence qu'une part significative des gènes organisés en îlot présentent des profils d'expression similaires et/ou ont la même fonction et/ou interviennent dans le même processus biologique. De plus, à l'échelle du chromosome 3B entier, des mécanismes de régulation à longue distance entre îlots de gènes ont été suspectés.

En conclusion, cette étude a permis pour la première fois de mettre en évidence des relations entre la structure du génome, l'évolution, la fonction et la régulation des gènes à l'échelle d'un chromosome de blé. Le séquençage et l'annotation du chromosome 3B ainsi que l'utilisation de technologies telles que le RNAseq permettront d'analyser ces relations de façon encore plus précise et exhaustive.

Mots clés : Blé, structure du génome, espace génique, carte transcriptionnelle, évolution des gènes, fonction des gènes

Abstract

Because of its size (17 Gb), allohexaploid nature and high repeat content (>80%), the bread wheat genome has always been perceived as too complex for efficient molecular studies. As a consequence, our knowledge of the wheat genome structure is still limited. Following a chromosome-specific approach, the physical map of wheat chromosome 3B has recently been constructed and allowed the development of unique genomic resources. During my PhD the use of transcriptomic approaches based on these resources allowed me analysing the relationships between the structure of the genome, the evolution, the function and the regulation of the genes along wheat chromosome 3B.

First macroarrays carrying the BACs of the chromosome 3B "Minimal Tiling Path" (MTP) were hybridised with 15 mRNA samples to identify the BACs carrying genes. Then barley Agilent 15K expression microarrays were hybridised with the MTP of chromosome 3B pooled in three-dimension (3D) to precisely locate the genes on the physical map. To build the first transcription map of a wheat chromosome, the 3D pools as well as the 15 mRNA samples were hybridised onto wheat NimbleGen 40K expression microarrays. The results from these experiments allowed drawing some conclusions about gene space organisation on chromosome 3B. Thus the genes are spread all along chromosome 3B with a gradient of the gene density from the centromere to the telomeres with a higher proportion of genes organised in islands at the telomeres. An evolutionary analysis demonstrated that the islands would essentially be composed of genes that have undergone rearrangements in the wheat genome. Furthermore the transcription map also showed that a significant fraction of the genes organised in islands display similar expression profiles and/or share the same function and/or play a role in the same biological process. Moreover, at the scale of the whole chromosome 3B, mechanisms of long distance regulation between gene islands were suspected.

In conclusion this study allowed for the first time to find relationships between the genome structure, the evolution, the function and the regulation of the genes at a wheat chromosome scale. The sequencing and the annotation of chromosome 3B as well as the use of technologies like RNAseq will enable to analyse these relationships in an even more precise and exhaustive way.

Keywords: Wheat, genome structure, gene space, transcription map, evolution of genes, function of genes

Remerciements

Après presque quatre années passées dans l'équipe « Structure, Fonction et Evolution des génomes de blé », les personnes qui me connaissent savent à quel point les remerciements représentent une partie difficile à écrire pour moi. Non pas que j'ai du mal à exprimer mes sentiments mais plutôt que mes yeux emplis de larmes compliquent considérablement la rédaction de toute ma gratitude. Ces larmes représentent toute l'émotion que j'ai en repensant à toute la passion, la bonne humeur, les éclats de rire partagés pendant cette thèse mais aussi au soutien et à l'encouragement que chacun a exprimé à mon égard dans les moments plus difficiles. Ainsi je voudrais remercier...

La Région Auvergne pour avoir financé ce projet de thèse mais aussi Gilles Charmet et Michel Bernard pour m'avoir accueillie dans l'unité.

Catherine Feuillet, cette femme en or qui malgré la charge de travail considérable et les nombreux déplacements sait rester à l'écoute de ses petits « poussins » avec une disponibilité et une simplicité exemplaires. Merci de m'avoir accueillie dans la famille Génome. Merci aussi pour ta confiance et ton soutien sans failles mais aussi pour ton dynamisme et ta bonne humeur légendaires. J'aimerai pouvoir un jour te ressembler ne serait-ce qu'un tout petit peu...

Etienne Paux pour son aide, son soutien, sa confiance, ses encouragements, sa disponibilité et sa bonne humeur au quotidien. Tu as su me transmettre ta passion et me faire grandir pendant ces quatre années. Comme on apprend à un enfant à faire du vélo sans petites roues, tu m'as permis de prendre mon envol. Tu as donné sans compter pour que mon travail soit mis en valeur le plus possible. Pour avoir été bien plus que l'encadrant de ma thèse et comme les mots me manquent, je te dis simplement merci infiniment.

Christel Laugier pour avoir été l'amie et la confidente de tous les jours aussi bien dans les bons que dans les mauvais moments. Merci pour ces milliers d'éclats de rire et moments de délire mais aussi pour ta présence les jours où le moral ne suivait pas.

Frédéric Choulet pour ton aide et ton soutien mais aussi pour les nombreuses discussions, la pertinence et la justesse de tes réflexions toujours constructives.

Pierre Sourdille pour les discussions, les réflexions et toutes ces fois où nous avons refait le monde. Merci aussi pour ton soutien au quotidien.

Alicia Besson, Nelly Cubizolles, Emmanuelle Lagendijk et Marion Ranoux pour leurs encouragements, leur bonne humeur et leur amitié.

Bouzid Charef, Jacqueline Philippon, Delphine Boyer, Isabelle Bertin, Nelly Cubizolles et Jérémy Martin pour leur aide technique précieuse, leur serviabilité et leur bonne humeur.

Romain Philippe et Arnaud Remay, en tant que jeunes docteurs, pour avoir su trouver les mots justes de soutien et d'encouragement au moment de la rédaction de cette thèse.

Le groupe du midi pour les discussions enflammées à la cantine et les plus de 2000 parties de tarot jouées pendant ces quatre années.

Le groupe du mercredi soir pour les soirées de détente et d'amusement devant une bonne bière.

L'équipe Génome pour sa cohésion, son implication, son dynamisme, son sérieux et son esprit de famille. Merci à tous pour votre soutien. Je voudrais aussi remercier ceux qui ont quitté l'équipe comme Cyrille Saintenac, Walid Alfares ou Sébastien Faure qui font toujours partie de la famille.

Ma famille qui malgré la distance, m'a toujours soutenue et a toujours cru en moi et en mes choix. Je voudrais leur dire que c'est leur amour, leur confiance et leur soutien qui m'a permis d'arriver jusqu'ici. Merci infiniment, je vous aime tous très fort.

Patrick, qui, malgré cette difficile période de rédaction, est toujours décidé à me dire « Oui ». Une grande partie du mérite de ce travail te revient. Tu as toujours su trouver les mots, les gestes et les attentions dont j'avais besoin pendant les périodes de doute, de frustration et de déception. Tu as toujours encouragé mes choix malgré les contraintes qu'ils ont engendrées dans notre vie de couple. Cette thèse, comme moi-même, ne serait qu'une pâle copie de ce qu'elle est aujourd'hui si tu n'avais pas été à mes côtés.

Toutes les personnes qui ont participé de près ou de loin à ce travail.

Abréviations

3C : Chromosome Conformation Capture
3D : tridimensionnel
ADN : Acide DésoxyriboNucléique
ADNc : ADN complémentaire
ANR : Agence Nationale pour la Recherche
ARN : Acide RiboNucléique
BAC : Bacterial Artificial Chromosome
bp : base pair
cDNA : complementary DNA
CGH : Comparative Genomic Hybridization
ChIP : Chromatin ImmunoPrecipitation
ChIP-chip : ChIP sur puce
ChIP–seq : ChIP par séquençage
cis-NAT : cis-Natural Antisens Transcript
CNV : Copy Number Variation
DNA : DeoxyriboNucleic Acid
DSB : Double Strand Break
EST : Expressed Sequence Tag
ET : Elément Transposable
FISH : Fluorescent In Situ Hybridization
3D-FISH: FISH tridimensionnel
FPC : Fingerprinted Contigs Program
Gb : Giga paires de bases
GFP : Green Fluorescent Protein
GO : Gene Ontology
HICF : High Information Content Fingerprinting
ISBP : Insertion Site-Based Polymorphism
kb : kilo paires de bases
LARD : LArge Retrotransposon Derivative
LINE : Long Interspersed Nuclear Element
LTR : Long Terminal Repeat
Mb : Mega paires de bases
MDR : Mathematically Defined Repeat
MeDIP : Methylation DNA ImmunoPrecipitation
MeDIP-chip : MeDIP sur puce
MeDIP–seq : MeDIP par séquençage
MITE : Miniatures Inverted-repeat Transposable Elements
MTP : Minimal Tiling Path
NCBI : National Center for Biotechnology Information
nt : nucléotides
PAV : Presence/Absence Variation
pb : paires de bases
PCR : Polymerase Chain Reaction
pg : picogramme
Q : Questionable
Q-PCR : PCR quantitative
QRT-PCR : Reverse Transcription Q-PCR
RNA : RiboNucleic Acid
SINE : Short Interspersed Nuclear Elements
siRNA : small interfering RNA
SNP : Single Nucleotide Polymorphism
SSR : Simple Sequence Repeat
TE : Transposable Element
TRIM : Terminal Repeats In Miniature

SOMMAIRE

Avant-propos

Le blé tendre, *Triticum aestivum* L., est une monoctylédone de la famille des Poacées au même titre que d'autres graminées cultivées, comme l'orge, le riz, le maïs et le sorgho. Le blé fait partie de la sous-famille des Pooideae qui contient également l'orge, le seigle et l'avoine. Mais le blé tendre se distingue des autres espèces de Pooideae par son importance socio-économique majeure puisqu'il représente la deuxième source de calories par habitant et par an dans le monde après le riz en 2007 (FAOSTAT) et représente la nourriture de base pour 35% de la population mondiale (Williams, 1993). En 2009, le blé était la 1ère culture en termes de surface cultivée dans le monde avec plus de 225 millions d'hectares et la 3ème en termes de production avec presque 682 millions de tonnes produites, derrière la canne à sucre et le maïs (FAOSTAT). La production de blé est très importante au niveau européen et français puisque l'Europe est le premier producteur mondial et la France est le premier producteur européen.

Cependant l'agriculture mondiale se trouve actuellement face à trois défis majeurs : i) répondre à la demande alimentaire changeante d'une population mondiale grandissante, ii) le faire dans des conditions environnementales et sociales durables et iii) assurer que les populations les plus pauvres ne vivent plus dans la faim (Godfray et al., 2010). En effet, l'utilisation des ressources alimentaires se diversifie notamment du fait de l'émergence des biocarburants (Godfray et al., 2010). De plus, la population mondiale devrait atteindre un plateau aux environs des 9 milliards d'habitants d'ici le milieu du siècle et l'objectif serait de produire 70% d'alimentation en plus dans le même délai (Godfray et al., 2010; Tester and Langridge, 2010). Des progrès sans précédent depuis la Révolution Verte des années 60 doivent être donc réalisés au niveau de l'amélioration variétale et des pratiques agronomiques pour permettre cette augmentation de la production alimentaire (Evans, 2009; Tester and Langridge, 2010). Parvenir à cette augmentation serait envisageable dans un environnement stable mais paraît plus douteux dans un environnement soumis au changement climatique (Tester and Langridge, 2010).

Différents facteurs importants pour les productions végétales sont, en effet, influencés par le changement climatique. Le CO_2 dont la teneur augmente dans l'air atmosphérique, agirait comme un engrais sur les céréales avec un métabolisme en C3, comme le blé ou le riz.

Cependant cet effet positif serait contrebalancé par la hausse des températures ainsi que la sécheresse (Evans, 2009). De plus, l'augmentation de la fréquence des phénomènes climatiques extrêmes, de la teneur en ozone ou l'émergence de nouvelles maladies des cultures sont autant de facteurs néfastes difficiles à quantifier. Ainsi, l'évaluation de l'impact réel du changement climatique sur les cultures reste incertaine mais anticiper ces effets néfastes par l'amélioration de la résistance des plantes aux stress abiotiques tels que la sécheresse ou la salinité parait indispensable (Tester and Langridge, 2010).

Ainsi, la production de denrées alimentaires dont le blé représente une part importante, doit être considérablement optimisée et dans un temps très limité. D'après Tester et Langridge (2010), la réussite de ce défi que l'humanité s'apprête à relever, implique notamment une meilleure connaissance des génomes des plantes, une meilleure compréhension des caractères phénotypiques, une meilleure maîtrise des outils de génomiques et de biotechnologies et une meilleure intégration de ces derniers dans les programmes de sélection.

Comme présenté dans la suite de ce document, ce défi pourrait s'avérer être encore plus laborieux à relever pour la communauté scientifique du blé tendre puisque la connaissance de son génome et les programmes de sélection qui lui sont consacrés sont freinés par la complexité de la structure de son génome.

SYNTHESE BIBLIOGRAPHIQUE

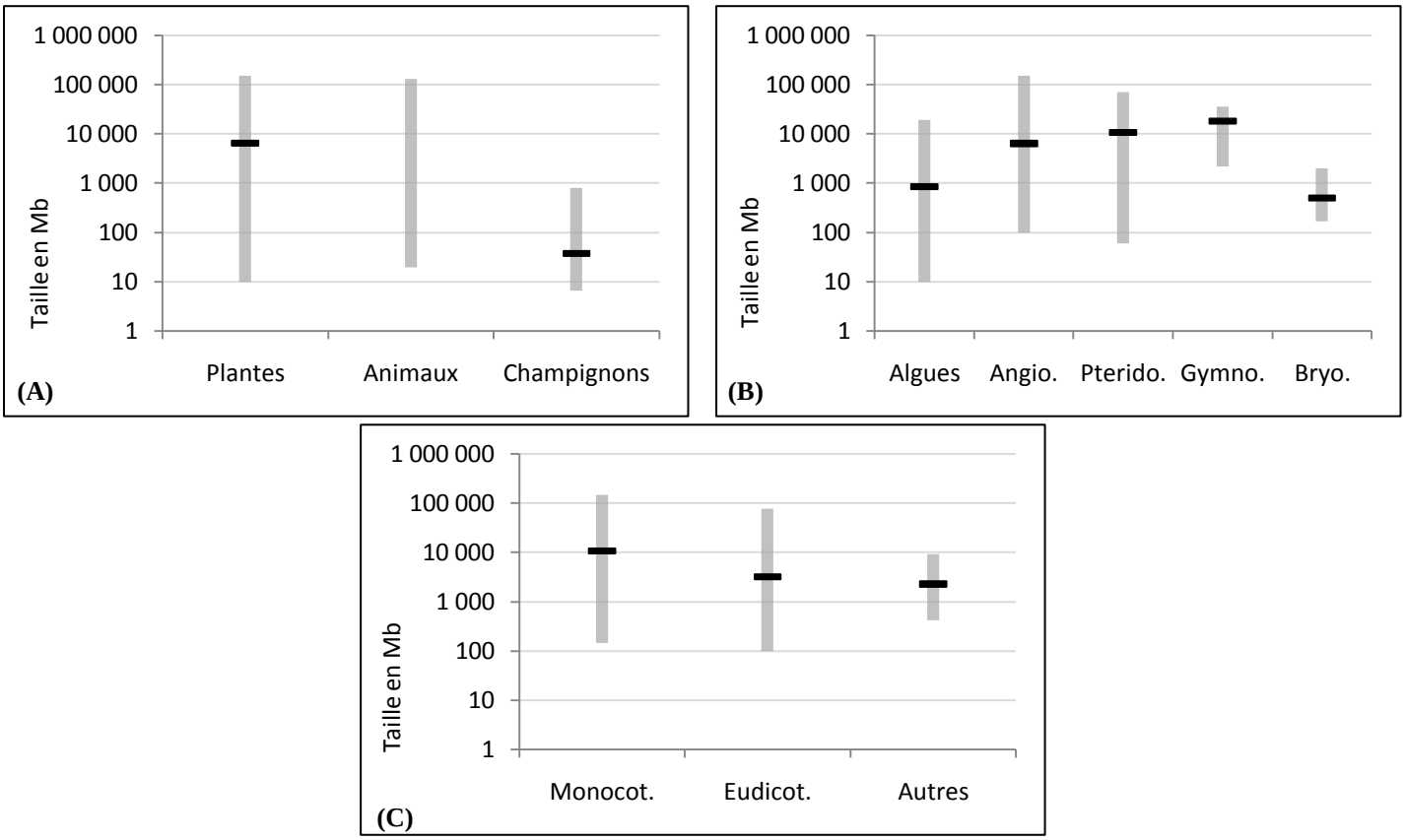


Figure 1 : Les variations de taille de génomes. Les bandes grises représentent la gamme de taille des génomes sur une échelle logarithmique et les traits noirs la taille moyenne par catégories. (A) Les variations de taille de génomes des eucaryotes les plus étudiés. La taille moyenne des génomes des animaux n'est pas fournie dans la base de données. (B) Les variations de taille de génomes de plantes. (C) Les variations de taille de génomes des angiospermes. (d'après Kullman et al., 2005; Animal genome size database; Plant DNA C-values database).

1. DES TAILLES DE GENOME TRES VARIABLES CHEZ LES PLANTES

1.1. Quelques chiffres

Depuis plus de cinquante ans, la taille des génomes est estimée grâce à la « valeur C » d'un organisme qui représente la quantité d'ADN nucléaire en picogrammes présente dans le noyau des gamètes à l'état non répliqué (Swift, 1950; Greilhuber et al., 2005; Plant DNA C-values database). Cette valeur est directement convertible en nombre de paires de bases puisque 1 pg = 978 Mb (Dolezel et al., 2003). Avec l'amélioration des techniques de quantification, la valeur C de milliers d'organismes a pu être déterminée (Animal genome size database; Kullman et al., 2005; Plant DNA C-values database). Pour les procaryotes, la quantification de l'ADN est réalisée directement en paires de bases par des méthodes plus adaptées aux génomes de très petite taille comme l'électrophorèse en champ pulsé ou le séquençage complet du génome (Prokaryote Genome Size Database).

La taille des génomes eucaryotes les plus étudiés (champignons, animaux et végétaux) est extrêmement variable puisque la gamme s'étend de 6,5 Mb pour le champignon *Pneumocystis carinii* f. sp. *muris* (Kullman et al., 2005) à presque 150 Gb pour la monocotylédone *Paris japonica* (Pellicer et al., 2010) soit une variation de 23000 fois.

Parmi les trois règnes, les champignons possèdent les génomes les moins variables en taille puisque 90% des génomes ont une taille entre 10 et 60 Mb avec une médiane de 28 Mb (Kullman et al., 2005; Gregory et al., 2007). Les animaux ont des génomes beaucoup plus variables avec des tailles extrêmes différant de plus de 6500 fois (Animal genome size database). C'est cependant le règne végétal qui présente les tailles de génomes les plus variables avec une gamme allant de 10 Mb pour l'algue *Ostreococcus tauri* à presque 150 Gb pour la monocotylédone *Paris japonica* soit une variation de 15000 fois avec une moyenne d'environ 6,5 Gb (Figure 1A) (Pellicer et al., 2010; Plant DNA C-values database).

Plus particulièrement parmi les plantes (Figure 1B), le groupe des angiospermes compte parmi les génomes les plus variables en taille avec une variation de presque 1500 fois entre les deux génomes extrêmes. Pour comparaison, le groupe des algues comprend des génomes extrêmes variant d'environ 2000 fois en taille, les ptéridophytes de 1200 fois, les

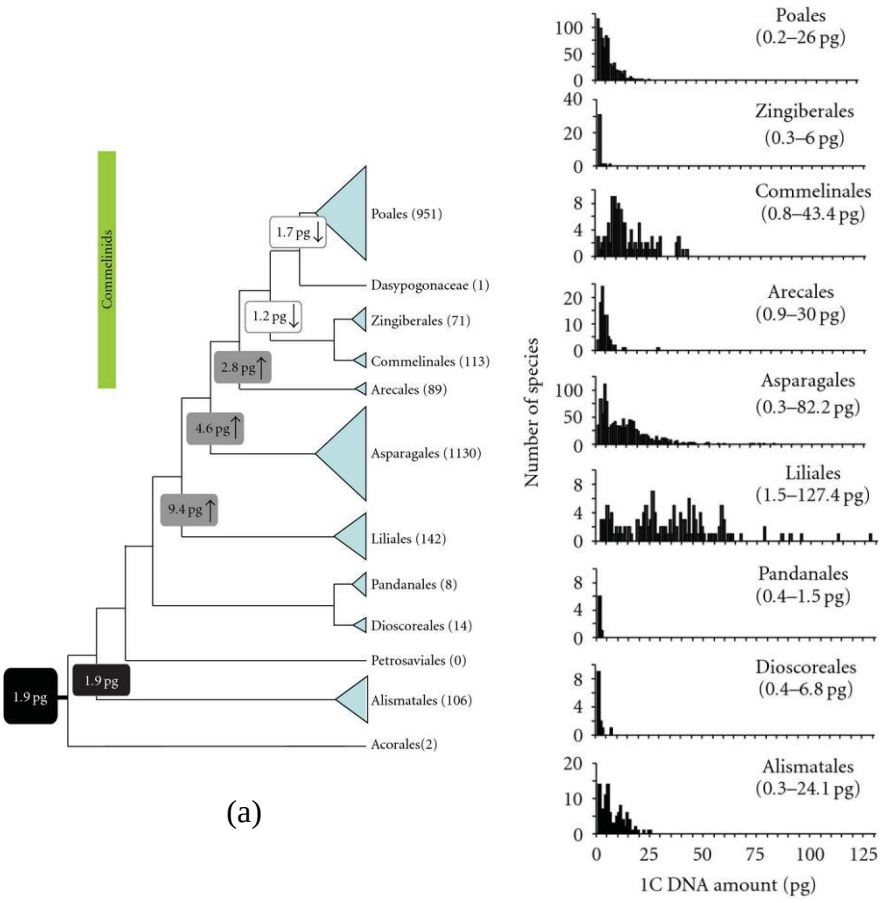


Figure 2 : La distribution phylogénétique de la taille des génomes des Monocotylédones. (a) Le résumé de la topologie des Monocotylédones avec entre parenthèses le nombre d'espèces avec des données de taille de génome. La taille du génome ancestral est spécifiée au niveau de différents nœuds. Les flèches indiquent le sens de variation de la taille du génome par rapport au génome ancêtre à la base de toutes les Monocotylédones. (b) Des histogrammes montrant la distribution des tailles de génome au sein des différents ordres avec les gammes de variation indiquées entre parenthèses. (d'après Leitch et al., 2010)

gymnospermes de 16 fois et les bryophytes de 12 fois. Au sein du groupe des angiospermes, les monocotylédones présentent une variation de taille de génome de l'ordre de 1000 fois et les eudicotylédones d'environ 800 fois (Figure 1C) (Pellicer et al., 2010; Plant DNA C-values database). De plus, parmi les angiospermes, des variations de taille de génome sont également observées au sein des genres avec une moyenne de variation de trois fois et un maximum observé pour le genre *Gossypium* de plus de 63 fois (Hawkins et al., 2008). Enfin des variations sont également observables au sein même des espèces comme pour *Hordeum spontaneum* (Kalendar et al., 2000).

Récemment Leitch et al. (2010) ont estimé que le génome de l'ancêtre des monocotylédones était petit et comptait 1,8 Gb (Figure 2). Au cours de l'évolution des monocotylédones, la tendance générale a été à l'augmentation de la taille des génomes à l'exception de certains ordres comme les Poales pour lesquelles la taille du génome ancêtre a légèrement diminué (1,7 Gb). Une telle alternance entre les augmentations et les diminutions de taille des génomes a également été observée dans d'autres familles de plantes et semble être la dynamique générale dans les génomes (Leitch et al., 2005). De nombreux mécanismes différents impliqués dans les variations de taille des génomes ont été décrits.

1.2. La polyploïdie, l'alternance entre l'augmentation et la diminution de la taille des génomes

De nombreuses études considèrent la polyploïdie comme un mécanisme majeur responsable de l'augmentation considérable et rapide de la taille des génomes (Adams and Wendel, 2005; Bennetzen et al., 2005; Comai, 2005; Gregory, 2005; Leitch and Leitch, 2008; Grover and Wendel, 2010). La polyploïdie est l'état héritable d'une cellule ou d'un organisme possédant un multiple supérieur à deux de lots haploïdes de chromosomes (Comai, 2005).

Les polyploïdes sont divisés en deux catégories selon la composition de leur génome et leur mécanisme de formation. Les autopolyploïdes possèdent des lots de chromosomes multiples ayant la même origine. Ils sont formés suite à des anomalies méiotiques rares générant des gamètes non réduits, c'est-à-dire possédant le stock diploïde de chromosomes. Ces gamètes non réduits fusionnent généralement avec des gamètes haploïdes pour produire un individu triploïde instable qui peut contribuer à la formation d'un individu tétraploïde plus stable

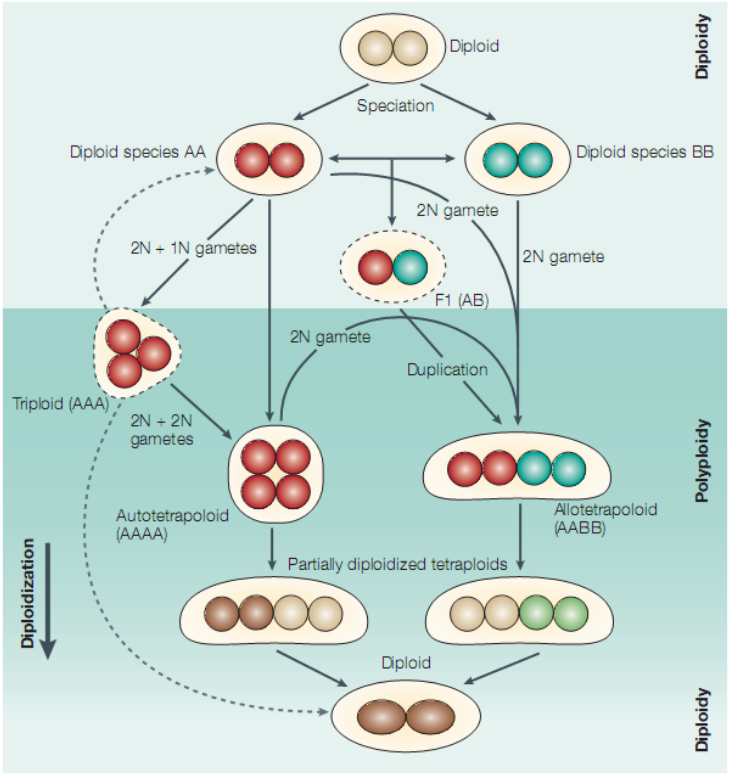


Figure 3 : L'alternance évolutive entre la diploïdie et la polyploïdie. La figure présente les différentes voies possibles de transition brutale de la diploïdie à la polyploïdie et de transition graduelle de la polyploïdie à la diploïdie. Les génomes haploïdes sont représentés par des disques colorés ou des ovales dans des noyaux beiges. Les génomes illustrés par des formes ovales représentent l'augmentation du nombre de gènes après rétention et subfonctionnalisation des gènes pendant la diploïdisation. Les disques ou ovales de couleurs différentes représentent des génomes ayant divergés. Les formes de ploïdie les plus instables ont des contours nucléaires en pointillés. (d'après Comai, 2005)

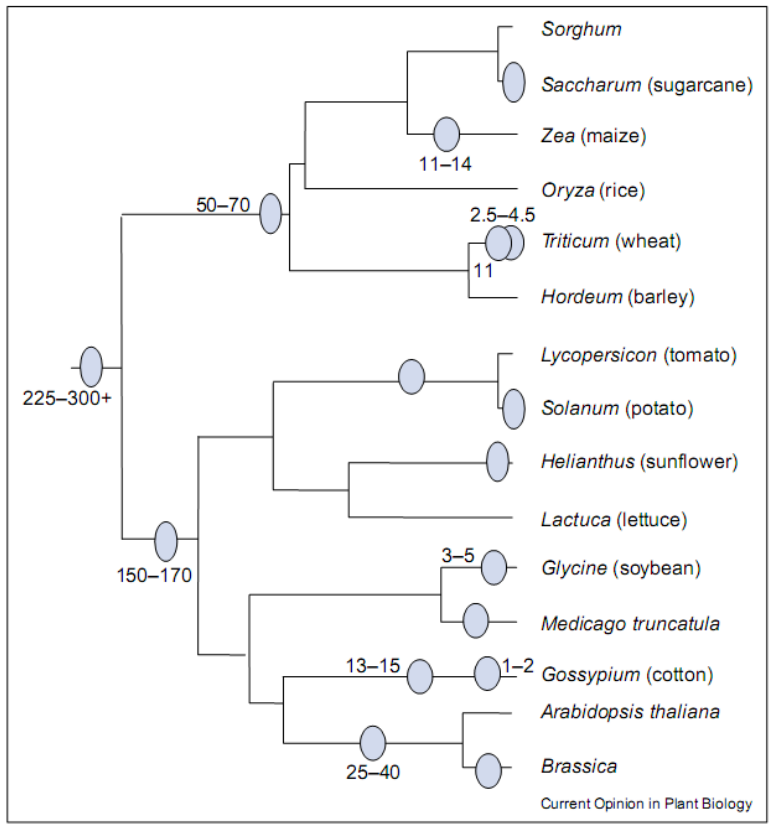


Figure 4 : Déduction des évènements de polyploïdie pendant l'évolution des Angiospermes. Les ovales bleutés représentent les évènements de polyploïdisation présumés et les nombres indiquent le temps approximativement écoulé (en millions d'années) depuis ces évènements. (d'après Adams et Wendel, 2005)

comme schématisé sur la figure 3 (Comai, 2005). Les allopolyploïdes, quant à eux, possèdent des lots de chromosomes multiples mais d'origines différentes. Ils résultent soit d'un hybride inter-spécifique instable ayant subi un doublement chromosomique, soit de la fusion de deux gamètes non réduits issus de deux espèces différentes comme schématisé sur la figure 3 (Comai, 2005).

Chez les plantes à fleurs, la fréquence d'apparition de polyploïdes est de l'ordre de un par 100000 et est bien plus importante que chez les animaux (Comai, 2005; Gregory, 2005; Leitch and Leitch, 2008). En effet, la formation de gamètes non réduits semble d'une part plus importante chez les plantes mais surtout les triploïdes végétaux présentent de meilleurs taux de survie par rapport aux animaux (Leitch and Leitch, 2008). Ainsi à l'heure actuelle de nombreuses espèces végétales et notamment un grand nombre d'espèces cultivées sont des polyploïdes comme la pomme de terre, le colza, le blé, le coton, la canne à sucre ou la banane (Leitch and Leitch, 2008). De plus, les analyses moléculaires suggèrent que la quasi-totalité des angiospermes présentent des traces d'évènements de polyploïdie anciens et parfois successifs. Par exemple, *Arabidopsis* et d'autres espèces auraient subi au moins deux épisodes de polyploïdisation au cours de leur évolution (Adams and Wendel, 2005; Abrouk et al., 2010). Les ancêtres des monocotylédones et des eudicotylédones seraient également des polyploïdes comme présenté sur la figure 4 (Adams and Wendel, 2005).

La polyploïdie, du fait de la présence de multiples lots de chromosomes fortement ressemblants, peut induire des dysfonctionnements au moment de la mitose et de la méiose (Comai, 2005; Hufton and Panopoulou, 2009). De nombreux réarrangements structuraux pouvant impliqués de grands fragments de chromosomes ont ainsi été observés chez les polyploïdes et provoquent généralement des diminutions de la taille des génomes (Leitch and Leitch, 2008; Hufton and Panopoulou, 2009). Au fil du temps, ces réarrangements permettent aux génomes polyploïdes de devenir stables du point de vue de l'appariement chromosomique jusqu'à se comporter comme des génomes diploïdes. Ce processus, appelé diploïdisation, alterne ainsi avec la polyploïdisation au cours de l'évolution du génome des plantes (Doyle et al., 2008; Hufton and Panopoulou, 2009).

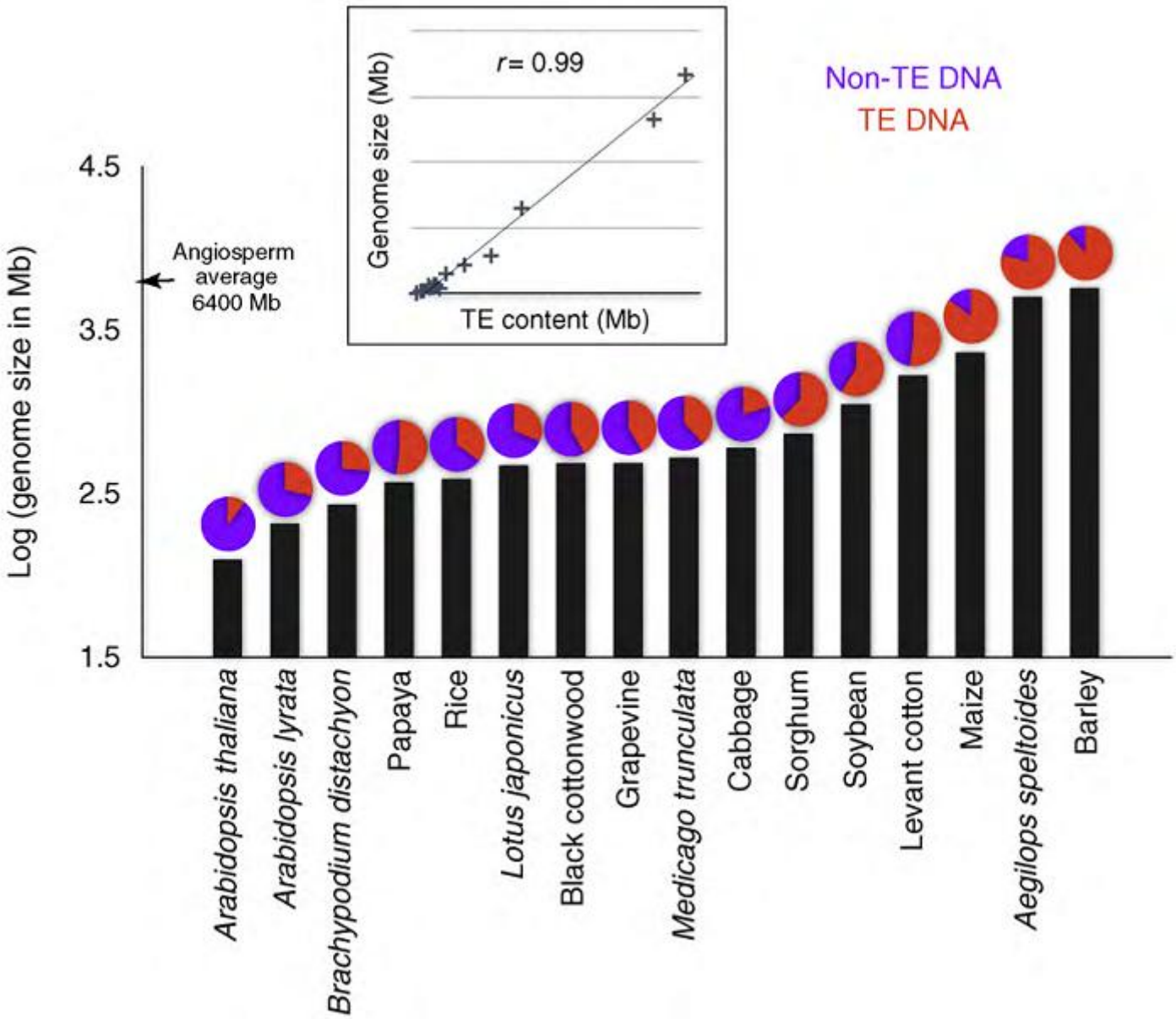


Figure 5 : Forte corrélation entre la taille du génome et la proportion d'ET chez les Angiospermes diploïdes. Les histogrammes représentent la taille des génomes (en base log) et les secteurs, les proportions relatives de séquences associées ou non aux ET. Le graphique en nuage de point représente la corrélation entre la taille des génomes et leur contenu en ET. (d'après Tenaillon et al., 2010)

1.3. Les séquences répétées, « l'aller simple vers l'obésité des génomes »

Dans les génomes, outre la polyploïdie, les séquences répétées, c'est-à-dire présentes en multiples copies, ont également un impact sur l'augmentation de la taille des génomes. Les séquences répétées sont de types variés et en proportions très variables. Celles-ci peuvent avoir des fonctions connues ou non. Les ARNr et les ARNt, impliqués dans la traduction, sont identifiés en grand nombre de copies dans les génomes (The *Arabidopsis* Genome Initiative, 2000; Tuskan et al., 2006; Jaillon et al., 2007; Merchant et al., 2007; Huang et al., 2009; Chan et al., 2010). Les séquences répétées télomériques qui stabilisent les extrémités des chromosomes, sont des répétions en tandem de séquences courtes du type CCCTAAA dans le génome d'*Arabidopsis* couvrant plusieurs kilobases (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Schmutz et al., 2010). Les séquences répétées centromériques fonctionnelles, impliquées dans l'attachement aux protéines du fuseau lors des divisions cellulaires, sont également constituées de répétitions en tandem de séquences entre 100 et 200 pb couvrant plusieurs dizaines de kilobases (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). D'autres séquences répétées en tandem généralement composées de motifs de une à quatre bases sont identifiées dans les génomes mais sans fonction associée et sont appelées microsatellites ou « Simple Sequence Repeats » (SSR) (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Merchant et al., 2007; Paterson et al., 2009; Chan et al., 2010; Schmutz et al., 2010).

Cependant, les principaux représentants parmi les séquences répétées sont les éléments transposables (ET), des éléments génétiques mobiles capables de se déplacer ou transposer, d'un emplacement à un autre au sein des génomes (Feschotte et al., 2002; Lisch, 2009). Les ET sont considérés comme les principaux responsables de l'augmentation de la taille des génomes chez les plantes puisque plus les génomes sont grands plus ils contiennent d'ET comme présenté sur la figure 5 (Tenaillon et al., 2010). Cependant différents types d'ET existent et sont présents en quantité variable dans les génomes. Les ET sont généralement classifiés en fonction de la nature de l'intermédiaire au moment de la transposition et en fonction de leur mécanisme de transposition (Sabot et al., 2004; Wicker et al., 2007).

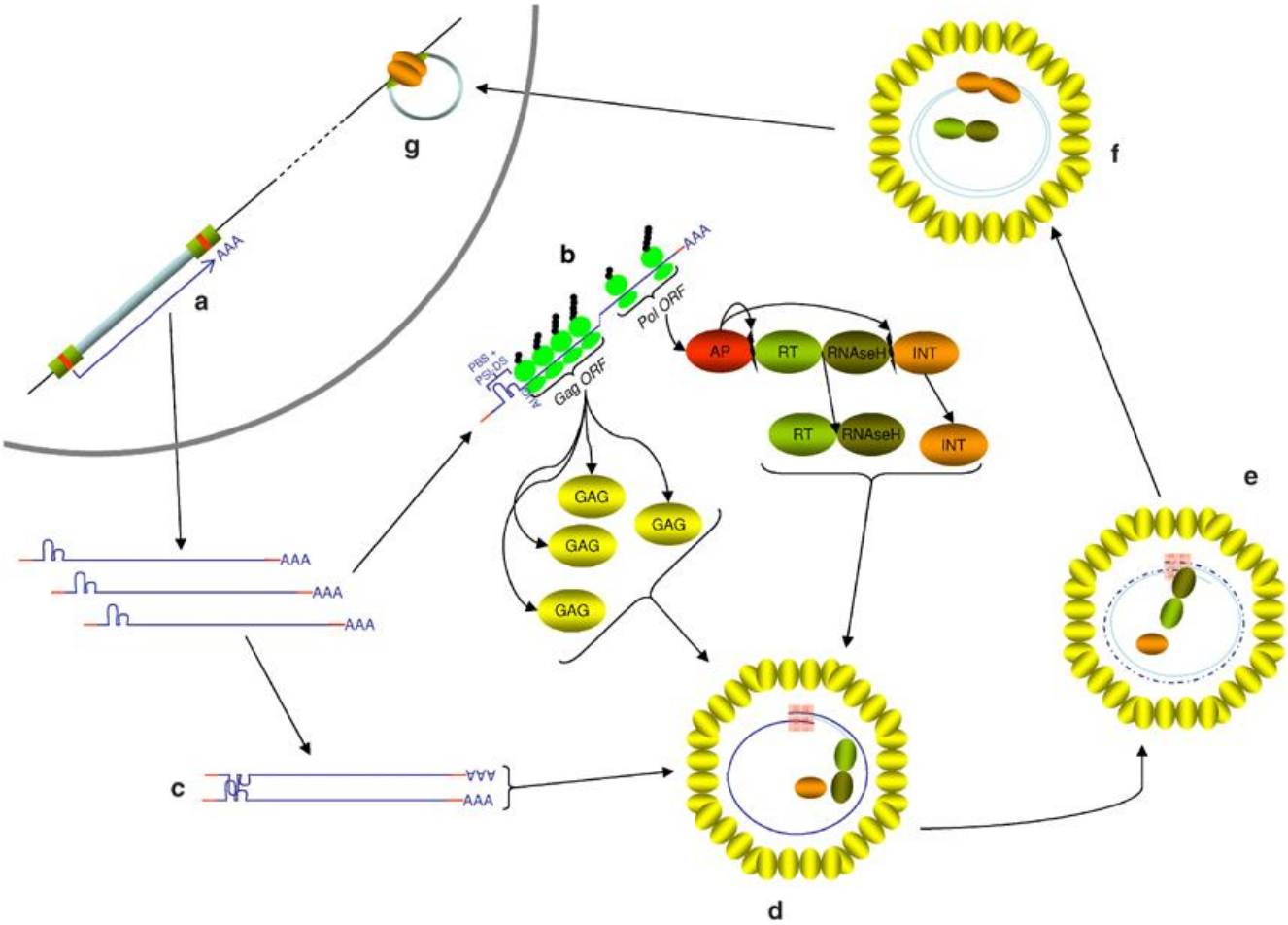


Figure 6 : Cycle de vie théorique des rétrotransposons à LTR. (a) Transcription de l'ARNm à partir des séquences régulatrices dans les LTR. (b) Traduction et synthèse protéique des éléments actifs GAG et POL ; la protéine POL est ensuite clivée par la protéinase AP pour libérer le complexe réverse-transcriptase-RNaseH (RT-RNAseH) et l'intégrase IN. (c) Dimérisation de l'ARN avant et pendant l'encapsidation. (d) Encapsidation de l'ARN et début de la réverse-transcription. Les protéines GAG polymérisent pour former une capside de type viral dans laquelle a lieu la réverse-transcription via le complexe RT-RNAseH. Ceci permet la synthèse du premier brin d'ADNc à partir de l'ARN encapsidé. (e) Dégradation de la matrice ARN et initiation de la synthèse du second brin d'ADNc. (f) Fin de la synthèse de l'ADNc double-brin et ligation de IN aux LTR. (g) Cassure double-brin et insertion de la copie néosynthétisée à un nouvel emplacement dans le génome. (d'après Sabot et Schulman, 2006)

1.3.1. Les différents types d'éléments transposables

1.3.1.1. Les éléments transposables de classe I

Les ET de classe I aussi appelés rétrotransposons transposent par transcription réverse d'un intermédiaire ARN (Kumar and Bennetzen, 1999; Bennetzen, 2000; Feschotte et al., 2002; Sabot et al., 2004; Wicker et al., 2007; Lisch, 2009). Pendant le transposition, cet intermédiaire ARN peut être directement réverse-transcrit à sa nouvelle position génomique ou passé par la formation d'un intermédiaire ADN par réverse-transcription dans le cytoplasme (Sabot et al., 2004; Wicker et al., 2007). La transposition des ET de classe I, du type « copier-coller », est décrite et schématisée sur la figure 6 (Sabot and Schulman, 2006). Pour les rétrotransposons dits autonomes, les protéines nécessaires à la transposition sont encodées par l'intermédiaire ARN lui-même (Feschotte et al., 2002; Sabot et al., 2004; Wicker et al., 2007). Cependant, les rétrotransposons dits non-autonomes qui restent capables de transposer, ne possèdent pas de séquences codantes mais ont conservé leurs séquences en *cis* indispensables à la transposition (Feschotte et al., 2002). Ils utilisent la machinerie d'ET autonomes pour transposer (Sabot and Schulman, 2006; Kapitonov and Jurka, 2007).

Les ET de classe I peuvent être divisés en deux sous-classes en fonction de leur mécanisme de transposition et de leur structure. Les rétrotransposons à « Long Terminal Repeat » (LTR) possèdent des séquences identiques, les LTR, en orientation directe à chacune de leurs extrémités d'une taille variant entre 100 pb et plusieurs kilobases (Bennetzen, 2000; Sabot et al., 2004). Les LTR sont notamment composés de séquences répétées régulatrices qui permettent d'initier et de terminer la transcription (Sabot and Schulman, 2006). Les rétrotransposons à LTR autonomes, dont les principaux représentants font partie des superfamilles *gypsy* et *copia*, codent au moins deux gènes, appelés *gag* et *pol* (Feschotte et al., 2002; Sabot et al., 2004). Le gène *gag* code une protéine de type capside et *pol* quatre protéines, une protéase, une réverse-transcriptase, une RNase H et une intégrase (Figure 6) (Feschotte et al., 2002; Sabot et al., 2004; Wicker et al., 2007). Les rétrotransposons à LTR non-autonomes, dont les principaux représentants font partie des superfamilles « LArge Retrotransposon Derivative » (LARD) et « Terminal Repeats In Miniature » (TRIM), ne possèdent pas ces séquences codantes entre leurs LTR (Sabot et al., 2004).

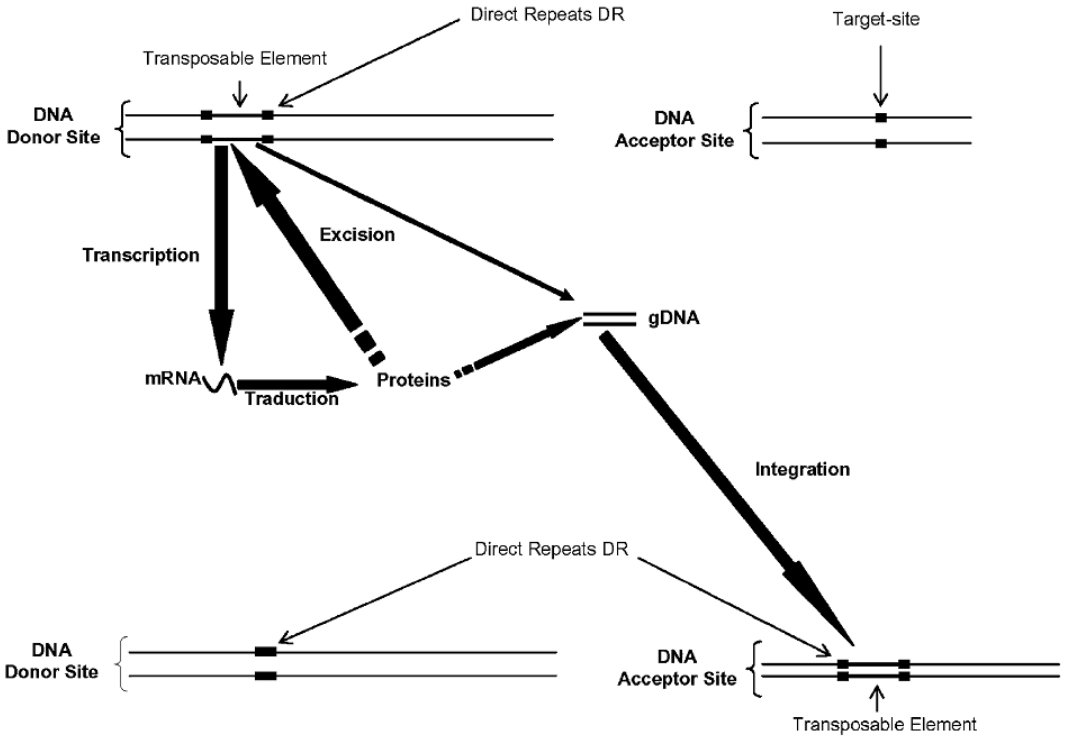


Figure 7 : Le système de transposition « couper-coller » des ET de classe II. Des séquences répétées ou « Direct Repeats » DR bordent l'ET. Après l'excision, les DR restent au niveau du site donneur laissant une trace de la transposition de l'ET. (d'après Sabot et al., 2004)

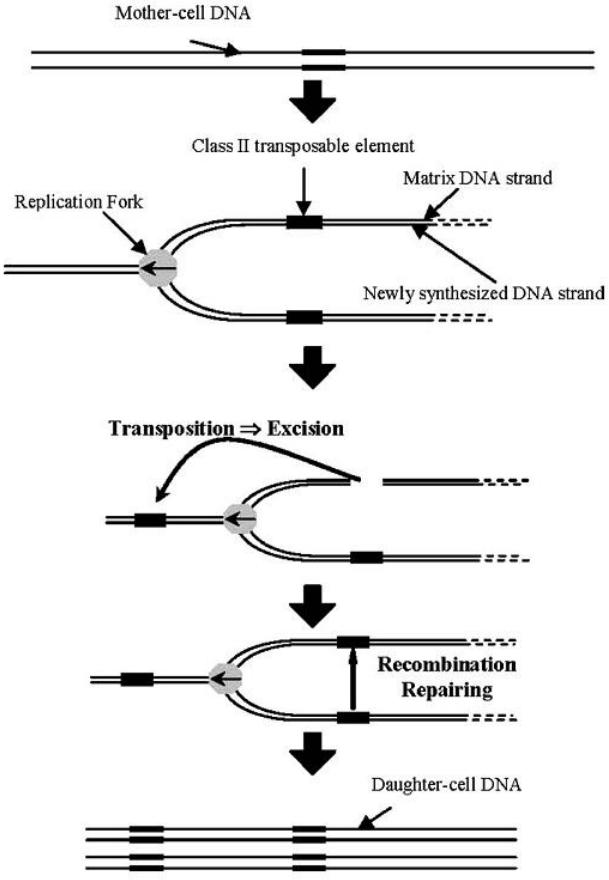


Figure 8 : Transposition des ET de classe II pendant la réplication de l'ADN en phase S du cycle cellulaire. La cassure double-brin générée lors de la transposition est réparée par recombinaison homologue avec la chromatide sœur. Au final, à partir d'un élément non réplicatif, deux copies sont créées. (d'après Sabot et al., 2004)

La deuxième sous-classe des ET de classe I sont les rétrotransposons sans LTR, dont les principaux représentants font partie des superfamilles « Long Interspersed Nuclear Elements » (LINE) et « Short Interspersed Nuclear Elements » (SINE). Les rétrotransposons sans LTR possèdent généralement une séquence poly(A) à leur extrémité 3'. Les LINE sont autonomes et codent des protéines similaires à *gag* et *pol* alors que les SINE sont non-autonomes mais possèdent des promoteurs pour l'ARN polymérase III (Feschotte et al., 2002; Sabot et al., 2004; Wicker et al., 2007).

1.3.1.2. Les éléments transposables de classe II

Les ET de classe II transposent via un intermédiaire ADN et de ce fait sont également appelés transposons à ADN (Feschotte et al., 2002; Sabot et al., 2004; Lisch, 2009). Les transposons à ADN transposent traditionnellement selon le mode « couper-coller » mais également selon le mode « copier-coller » dans certaines conditions (Figures 7 et 8) (Sabot et al., 2004; Wicker et al., 2007). La transposition s'opère par excision puis par intégration du fragment excisé via une transposase qui reconnait des sites spécifiques d'insertion (Sabot et al., 2004). Les ET de classe II sont tous bordés à leurs extrémités 5' et 3' par des séquences répétées de taille constante (quelques paires de base) et parfois même de séquence conservée au sein des superfamilles (Sabot et al., 2004). Les transposons à ADN autonomes codent le gène de la transposase alors que les non-autonomes ne possèdent pas de séquences codantes fonctionnelles (Feschotte et al., 2002; Sabot et al., 2004; Wicker et al., 2007; Lisch, 2009). Les principales superfamilles sont les CACTA, les hAT et les Mutator au sein desquelles des transposons autonomes et non-autonomes ont été identifiés (Feschotte et al., 2002; Sabot et al., 2004). Les « Miniatures Inverted-repeat Transposable Elements » (MITE) sont de très petits transposons à ADN non-autonomes (30 à 500 pb) puisqu'ils ne possèdent plus de séquences codantes (Sabot et al., 2004; Wicker et al., 2007).

Les *Helitrons* sont classés parmi les ET de classe II mais transposent selon un mécanisme réplicatif particulier dit « rolling-circle » décrit par Sabot et al. (2004). Ils possèdent une structure différente des autres ET de classe II avec peu de séquences caractéristiques qui puissent faciliter leur identification mais une séquence palindromique à leur extrémité 3' (Sabot et al., 2004; Lisch, 2009). Les *Helitrons* autonomes codent une nucléase-ligase et une

Tableau 1 :Les caractéristiques de quelques génomes de plantes séquencés

	Concom-bre	*Brachy-podium*	Ricin	Papaye	Riz	Peuplier	Vigne	Sorgho	Soja	Maïs	Orge
Taille en Mb	243,5	272	350,6	372	389	485	487	730	1115	2300	5700
Nombre de gènes	26682	25532	31237	24746	37544	45555	30434	34496	46430	32540	38000
%ET	24%	28,10%	50,30%	51,90%	34,80%	42%	41,40%	62%	59%	84,2%	85%
%classe1	12,16%	23,33%	18,16%	42,80%	19,35%	7,0%	27,69%	54,52%	42,24%	75,60%	67,61%
%LTR-RT*	10,40%	21,40%	16,22%	33,30%	14,75%	6,5%	23,20%	54,43%	42%	75%	66,99%
%classe2	1,24%	4,77%	0,91%	0,19%	12,96%	2,1%	0,84%	7,46%	16,50%	8,60%	6,44%
%autres ET	10,60%	0,00%	31,23%	8,91%	2,49%	32,84%	12,87%	0,02%	0,26%	0,00%	10,95%

* LTR-RT : rétrotransposon à LTR

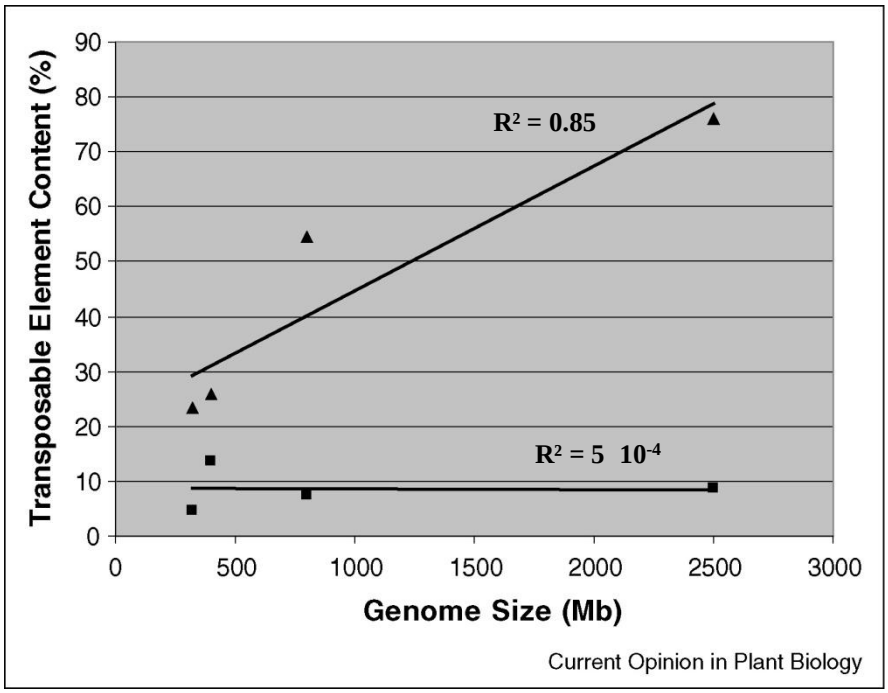


Figure 9 : Corrélation entre la taille des génomes et la proportion de rétrotransposons à LTR (triangles) ou de transposons à ADN (carrés). Les données des génomes de Brachypodium, du riz, du sorgho et du maïs ont été utilisées pour tracer ce graphique. (d'après Devos, 2010)

hélicase impliquées dans la transposition (Sabot et al., 2004; Wicker and Keller, 2007; Lisch, 2009).

1.3.2. Les rétrotransposons à LTR responsables de l'obésité des génomes végétaux

Comme suggéré dans de très nombreuses études, les ET joueraient un rôle majeur dans l'augmentation de la taille des génomes. Cependant les classes et familles d'ET impliquées seraient particulières en fonction des règnes. En effet, chez les Mammifères, les LINE et les SINE seraient responsables de l'augmentation de la taille des génomes alors que chez les nématodes, les transposons à ADN en seraient les principaux acteurs (Gregory, 2005). Par contre, chez la Drosophile comme chez les plantes, les rétrotransposons à LTR sont les séquences répétées qui joueraient un rôle majeur dans l'augmentation de taille des génomes (Bennetzen and Kellogg, 1997; Sanmiguel and Bennetzen, 1998; Kalendar et al., 2000; Sabot et al., 2004; Gregory, 2005; Vitte and Panaud, 2005; Rabinowicz and Bennetzen, 2006; Sabot and Schulman, 2006; Vitte et al., 2007; Devos, 2010).

Dans les génomes des plantes, parmi tous les ET, les rétrotransposons à LTR occupent en général la proportion la plus importante comme présenté dans le tableau 1 (International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Jaillon et al., 2007; Ming et al., 2008; Huang et al., 2009; Mayer et al., 2009; Paterson et al., 2009; Schnable et al., 2009; Chan et al., 2010; Schmutz et al., 2010; The International *Brachypodium* Initiative, 2010). Cette proportion est très variable en fonction des génomes présentés mais est corrélée positivement avec la taille des génomes contrairement à la proportion de transposons à ADN (Figure 9) (Devos, 2010). En effet, les rétrotransposons à LTR et les MITE sont les ET les plus nombreux dans les génomes végétaux avec jusqu'à 120000 et 100000 copies respectivement (Casacuberta and Santiago, 2003; Sabot et al., 2004; Vitte and Panaud, 2005; Devos, 2010). Dans le génome du riz, les MITE sont même deux fois plus nombreux que les rétrotransposons à LTR mais ces derniers représentent une fraction plus importante du génome du fait de leur grande taille, à savoir entre 2 et 18 kb (Sabot et al., 2004; International Rice Genome Sequencing Project, 2005; Vitte et al., 2007). Ainsi les rétrotransposons à LTR seraient les principaux responsables de l'obésité des génomes végétaux grâce à leur nombre et

à leur taille mais également du fait de leur mécanisme de transposition réplicatif permettant d'augmenter très rapidement leur nombre (Bennetzen and Kellogg, 1997; Vitte and Panaud, 2005). En effet, les rétrotransposons à LTR semblent proliférer massivement dans les génomes des plantes sur des périodes évolutives exceptionnellement courtes (Hawkins et al., 2008). Par exemple, SanMiguel et al. (1998) ont estimé que la taille du génome du maïs avait doublé en trois millions d'années du fait de l'accumulation rapide de rétrotransposons. De même, dans le génome du riz, des invasions massives de rétrotransposons à LTR datant de moins de 5 millions d'années ont été identifiées (Vitte and Panaud, 2005; Vitte et al., 2007).

Plus particulièrement, de récentes études suggèrent que l'augmentation de taille des génomes serait liée à l'accumulation de rétrotransposons à LTR issus d'un nombre limité de familles. En effet, dans les génomes du riz et du maïs, la majorité des familles de rétrotransposons à LTR ne compte que très peu de membres (Baucom et al., 2009a; Baucom et al., 2009b). Ainsi, seules 32% et 37% des familles compteraient trois membres ou plus dans les génomes du riz et du maïs respectivement (Baucom et al., 2009a; Baucom et al., 2009b). Un autre exemple est énoncé chez l'orge dans le génome duquel 14 familles d'ET (2 familles de CACTA et 12 familles de rétrotransposons à LTR dont 10 *gypsy* et 2 *copia*) représentent plus de 50% de la taille du génome (Wicker et al., 2009). De plus, en comparant l'abondance de quelques familles de rétrotransposons à LTR dans différents génomes d'angiospermes, Vitte et Bennetzen (2006) ont mis en évidence une corrélation entre le nombre estimé de représentants dans les familles les plus abondantes et la taille des génomes. Les grands génomes auraient plusieurs familles de plus de 40000 copies alors que les plus petits génomes étudiés auraient peu ou pas de familles de plus de 1000 copies. Cette corrélation a également été avancée par Baucom et al. (2009a) pour expliquer la différence de taille entre les génomes du riz et du maïs.

Cependant, dans certains génomes, l'obésité des génomes liée aux ET et plus particulièrement aux rétrotransposons à LTR est contrée par des mécanismes de délétion visant ces derniers de façon préférentielle.

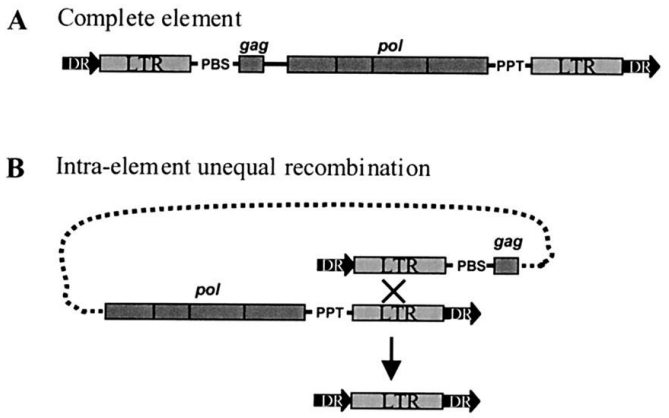


Figure 10 : Recombinaison homologue inégale entre rétrotransposons à LTR. (A) Structure d'un élément complet avec des « Direct Repeats » (DR) à chaque extrémité, deux LTR, un « Primer-Binding Site » (PBS), le « PolyPurine Tract » (PPT) nécessaire à la réplication de l'élément et les gènes (gag et pol). (B) Solo-LTR formé par recombinaison homologue au sein du même élément. Le ligne pointillée est utilisée pour faciliter la représentation du repliement nécessaire pour réaliser la recombinaison et ne représente une séquence d'ADN particulière. (d'après Devos et al., 2002)

1.4. La recombinaison, un mécanisme efficace pour réduire la taille des génomes ?

Des mécanismes de réduction contrant l'obésité des génomes végétaux ont été mis en évidence essentiellement par l'étude des rétrotransposons à LTR. Les deux mécanismes principalement identifiés sont la recombinaison illégitime et la recombinaison homologue inégale (Gregory, 2005; Vitte and Panaud, 2005; Hawkins et al., 2008; Grover and Wendel, 2010; Tenaillon et al., 2010). Ces deux mécanismes impliquent la reconnaissance de séquences homologues.

La recombinaison illégitime implique généralement la reconnaissance de séquences homologues courtes de l'ordre de 2 à 15 pb (Ma et al., 2004; Bennetzen et al., 2005). Une recombinaison illégitime peut se produire notamment lors d'erreurs pendant la réplication de l'ADN ou lors de réparations de cassures double-brin (Devos et al., 2002; Vitte and Panaud, 2005). Elle est la résultante d'une erreur dans la reconnaissance des séquences homologues qui peut générer un décalage et une perte d'ADN de quelques paires de base jusqu'à plusieurs centaines (Devos et al., 2002). Lorsqu'un évènement de recombinaison illégitime provoque une délétion dans un élément transposable, celui-ci apparaît alors tronqué (Devos et al., 2002; Tian et al., 2009).

Par contre, la recombinaison homologue inégale requiert des séquences homologues de taille plus importante (>50 pb) (Ma et al., 2004; Bennetzen et al., 2005). La trace d'un évènement de recombinaison homologue inégale est essentiellement visible lorsqu'il a lieu au sein d'un rétrotransposon à LTR ou entre plusieurs rétrotransposons à LTR. Comme schématisé sur la figure 10 (Devos et al., 2002), une recombinaison homologue inégale réalisée entre LTR génère essentiellement des structures du type solo-LTR résultant de la délétion du corps entier du rétrotransposon à LTR ainsi que d'un des deux LTR.

Les deux mécanismes de suppression d'ADN sont considérés comme plus ou moins efficaces en fonction des espèces. En effet, dans le génome d'*Arabidopsis*, la recombinaison illégitime aurait supprimé cinq fois plus d'ADN que la recombinaison homologue inégale (Devos et al., 2002). Dans le génome du riz par contre, la recombinaison homologue inégale serait cinq fois plus efficace que la recombinaison illégitime et la demi-vie des rétrotransposons à LTR serait de 3 à 6 millions d'années (Bennetzen et al., 2005; Vitte and Panaud, 2005; Vitte et al., 2007;

Tian et al., 2009). Au total le génome du riz aurait perdu plus de 190 Mb de séquence en quatre millions d'années (Ma et al., 2004).

Cependant l'efficacité de ces mécanismes pour contrer les invasions brutales des génomes par les ET reste controversée et semble très dépendante des espèces (Vitte and Bennetzen, 2006; Hawkins et al., 2008). De plus, cette efficacité peut également être régie par les conditions de vie des plantes puisque chez l'orge, Kalendar et al. (2000) ont constaté des adaptations de la taille des génomes en fonction de leur milieu de vie. Cette observation suggère ainsi un ordre supérieur dans la régulation de la taille des génomes.

1.5. L'épigénétique, une régulation de la taille des génomes ?

Des études de plus en plus nombreuses mettent en évidence le rôle majeur de la régulation épigénétique sur de nombreux mécanismes intervenant au sein des génomes. Ainsi les mécanismes impliqués dans l'évolution de la taille des génomes sont également régulés épigénétiquement.

Premièrement, entre les invasions brutales, la prolifération des ET est contenue grâce à des marques épigénétiques telles que le méthylation de l'ADN ou certaines modifications des histones qui permettent d'inactiver la transcription des régions concernées (Lisch, 2009; Grover and Wendel, 2010; Tenaillon et al., 2010). Par exemple, chez *Arabidopsis*, l'analyse de mutants du gène *ddm1* (« *decrease in DNA methylation 1* ») impliqué dans le remodelage de la chromatine, a mis en évidence la prolifération de certaines familles de rétrotransposons (Tsukahara et al., 2009). Mais la régulation de l'inactivation des ET semble fine et complexe car le relâchement d'une marque épigénétique particulière n'entraîne pas la réactivation de tous les ET mais seulement de quelques familles (Mirouze et al., 2009).

La mise en place des marques épigénétiques ciblant les ET des plantes, comme la méthylation de l'ADN ou les modifications d'histones, est réalisée par la voie des « small interfering RNA » (siRNA) à partir d'ARN issu d'ET transcrits et donc potentiellement actifs (Lisch, 2009; Grover and Wendel, 2010; Tenaillon et al., 2010). Ce mécanisme qui nécessite l'activité d'ET pour mettre en place leur inactivation est cohérente avec le fait que les invasions par les ET sont brutales et limitées dans le temps. Comme le nombre de transcrits

d'ET augmentent, le nombre de siRNA leur correspondant augmente ce qui favorise l'inactivation de l'élément actif par la mise en place des marques épigénétiques (Grover and Wendel, 2010). Une invasion résulterait d'un relâchement de marques épigénétiques potentiellement initié par des stress environnementaux (Kalendar et al., 2000) ou génomiques tels que l'hybridation interspécifique ou la polyploïdie (Lisch, 2009; Grover and Wendel, 2010; Tenaillon et al., 2010).

Deuxièmement, en plus de l'impact des marques épigénétiques sur la prolifération des ET et donc sur l'augmentation de la taille des génomes, celles-ci réguleraient également la diminution de la taille des génomes via la recombinaison. Ainsi, différentes études ont pu corréler les modifications de la structure de la chromatine avec les variations de la fréquence de recombinaison (Kantidze and Razin, 2009; Szekvolgyi and Nicolas, 2010). Ces modifications épigénétiques permettraient un contrôle de la recombinaison par l'environnement (Koren et al., 2002; Szekvolgyi and Nicolas, 2010).

En conclusion, la taille des génomes des plantes est très variable et est régie par de nombreux facteurs tels que la polyploïdie, la prolifération des éléments transposables, leur suppression via la recombinaison et probablement bien d'autres encore. Certains de ces facteurs sont régulés par des mécanismes épigénétiques fins et complexes encore peu connus jusqu'à présent. De plus, Grover et Wendel (2010) soulignent également le rôle majeur de la dynamique des populations avec des facteurs tels que la taille des populations, le système de culture ou la sélection qui contribuent au modelage des génomes.

2. L'ORGANISATION DE L'ESPACE GENIQUE CHEZ LES PLANTES

Dans ces génomes de taille très variables, des gènes sont bien entendu également présents. Dans cette seconde partie, la variabilité du contenu en gènes dans les génomes des plantes sera analysée afin de savoir si ce compartiment structural est aussi variable que ceux précédemment cités.

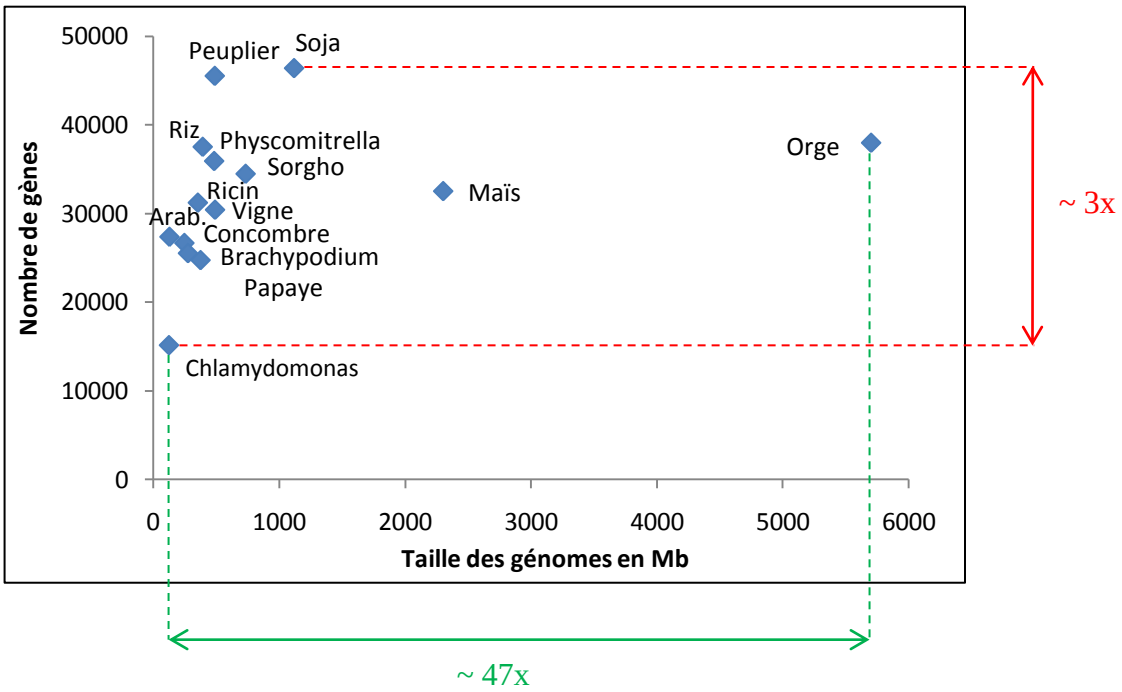


Figure 11 : Variation de la taille des génomes en fonction du nombre de gènes. La variation du nombre de gènes entre les génomes de plantes séquencés (en rouge) est d'environ trois fois alors que la variation de la taille des génomes (en vert) est d'environ 47 fois.

2.1. Le contenu génique bien conservé ou le paradoxe de la valeur C

2.1.1. Définition d'un gène

Le concept de « gène » n'a cessé d'évoluer et de se complexifier depuis son invention en 1909 par Wilhelm Johannsen pour lequel les gènes étaient à la fois des éléments permettant de spécifier les caractéristiques d'un organisme mais également les supports de l'hérédité de ces caractéristiques (Gerstein et al., 2007). Récemment, les études approfondies de cartographie de l'activité et de la régulation transcriptionnelles menées sur une fraction du génome humain au sein du projet ENCODE ont permis d'affiner la définition d'un gène. Ainsi comme défini suite au projet ENCODE, « le gène est une union de séquences génomiques codant un ensemble cohérent de produits fonctionnels potentiellement chevauchants » (Gerstein et al., 2007).

Dans le génome d'un organisme, on peut également qualifier la fraction du génome correspondant aux gènes en utilisant l'expression « espace génique ». Par extension l'espace génique peut également inclure la notion de distribution physique de ces gènes dans le génome (Jackson et al., 2004).

2.1.2. Quelques chiffres sur le contenu génique

Comme présenté sur la figure 11, le contenu en gènes entre les différentes espèces de plantes séquencées ou en cours de séquençage est relativement stable par rapport à la taille de leur génome (Devos, 2010). En effet leur contenu génique varie de 15000 gènes pour *Chlamydomonas* à 46000 gènes pour le soja soit une variation de trois fois alors que la taille des génomes s'étale entre 121 Mb pour *Chlamydomonas* à 5700 Mb pour l'orge soit une variation de 47 fois (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Jaillon et al., 2007; Merchant et al., 2007; Ming et al., 2008; Rensing et al., 2008; Huang et al., 2009; Mayer et al., 2009; Paterson et al., 2009; Schnable et al., 2009; Chan et al., 2010; Schmutz et al., 2010; The International *Brachypodium* Initiative, 2010). De plus, les génomes ayant les contenus en gènes les plus importants comme le soja ou le maïs sont généralement d'anciens polyploïdes qui auraient

conservés une certaine proportion de gènes dupliqués (Schnable et al., 2009; Schmutz et al., 2010). Ainsi, Devos (2010) estime que le nombre de gènes typiquement attendu dans un génome haploïde d'angiosperme serait proche de 28000.

D'après la figure 11, les génomes de grande taille comme ceux du maïs et de l'orge ont un contenu en gènes très appauvris par rapport à ce qu'on pourrait attendre avec des génomes de cette taille. Cette observation montre bien que la taille des génomes n'est pas corrélée à leur complexité en termes de nombre de gènes, à l'exception des génomes polyploïdes récents. Ce paradoxe est connu à la fois chez les plantes et chez les animaux sous le nom de paradoxe de la valeur C (Gregory, 2005). Ce paradoxe dit que la quantité d'ADN présente dans une cellule doit être constante puisque l'ADN est le constituant même des gènes et pourtant la quantité d'ADN n'est pas corrélée au nombre de gènes (Gregory, 2005).

De plus, la taille des gènes est également très stable entre les espèces végétales. En effet, la taille moyenne des exons multipliée par le nombre moyen d'exons par gène varie de 2,5 fois. De même, en incluant la taille moyenne des introns, la taille moyenne des gènes varie de 2,8 fois (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Jaillon et al., 2007; Merchant et al., 2007; Ming et al., 2008; Rensing et al., 2008; Huang et al., 2009; Paterson et al., 2009; Schnable et al., 2009; Chan et al., 2010; The International *Brachypodium* Initiative, 2010). Ainsi, aussi bien le nombre que la taille des gènes n'a pas d'influence sur la taille des génomes végétaux.

Finalement, la solution au paradoxe de la valeur C est désormais bien connue puisque la plupart des génomes sont constitués d'ET (Gregory, 2005). Ainsi, comme présenté précédemment, la taille des génomes est essentiellement corrélée avec la proportion de rétrotransposons à LTR chez les plantes alors que le contenu en gènes est stable.

2.1.3. La conservation des gènes entre espèces végétales

En plus de leur nombre et de leur taille, les gènes eux-mêmes semblent relativement bien conservés entre les génomes des plantes. De nombreuses études concernant la conservation des gènes entre espèces de plantes ont été menées et sont résumées notamment par Abrouk et al. (2010). Chez les eudicotylédones, 77% des gènes en moyenne sont conservés entre le

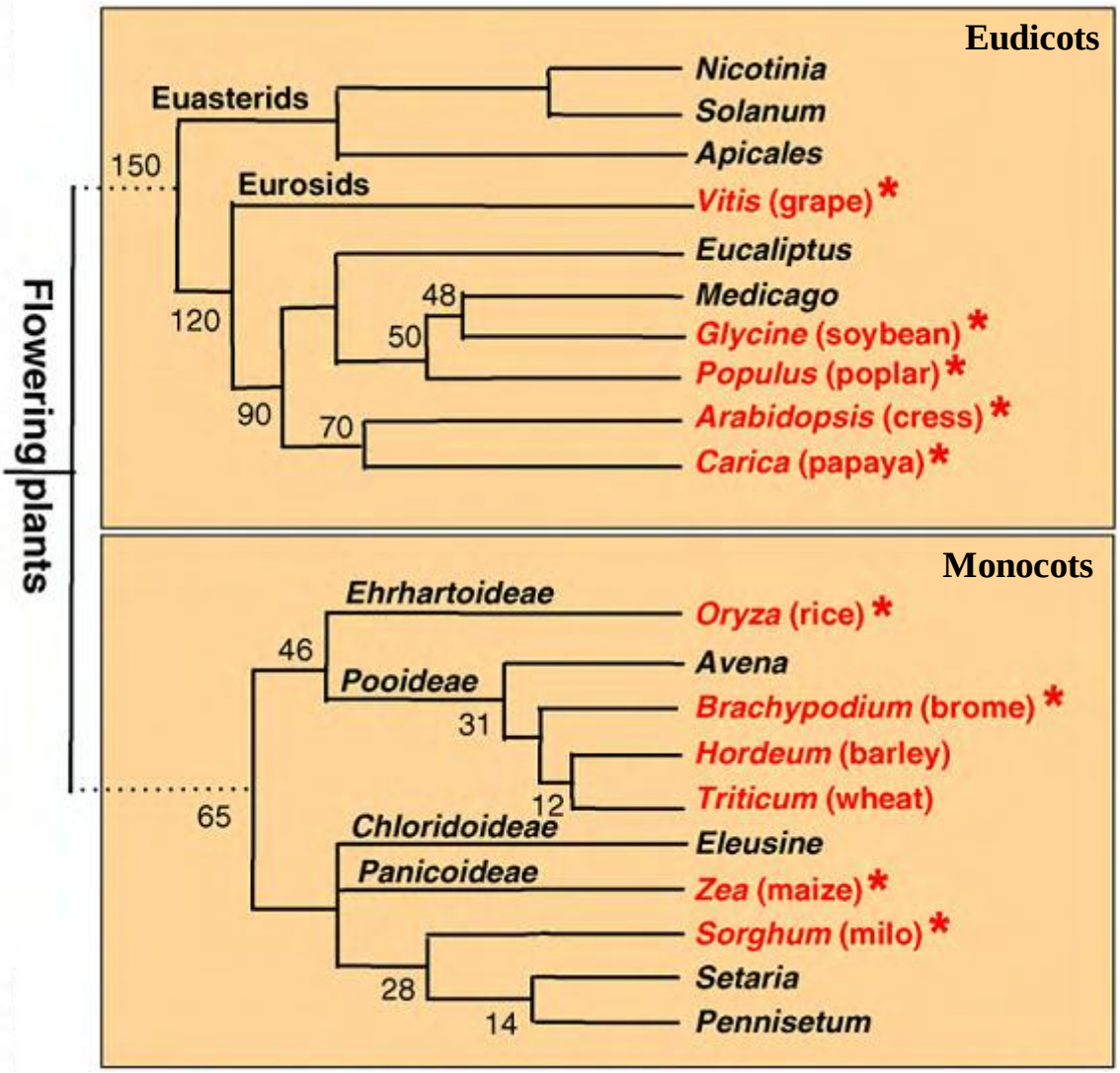


Figure 12 : Représentation schématique des relations phylogénétiques entre les espèces d'Angiosperme. Les temps de divergence (en millions d'années) à partir d'un ancêtre commun sont indiqués au niveau des branches de l'arbre phylogénétique. En rouge, les génomes utilisés par Abrouk et al. (2010) dans leur étude. Les génomes séquencés sont indiqués par des astérisques rouges. (d'après Abrouk et al., 2010)

génome de la vigne et les génomes d'*Arabidopsis*, du peuplier, du soja et de la papaye (Figure 12) (Abrouk et al., 2010). Chez les monocotylédones, entre 70 et 80% des gènes sont conservés entre les génomes en fonction des études et des espèces considérées parmi le riz, *Brachypodium*, le sorgho, le maïs, l'orge et le blé (Salse et al., 2008a; Bolot et al., 2009; Abrouk et al., 2010; Wicker et al., 2010). De plus, la position des gènes conservés entre ces différentes espèces a été analysée pour identifier la fraction conservée en position orthologue, c'est-à-dire en position identique par rapport à l'ancêtre commun. Seulement 17% et 16% des gènes en moyenne sont conservés en position orthologue entre la vigne et les quatre autres eudicotylédones étudiées et entre le riz et les quatre autres monocotylédones étudiées respectivement (Abrouk et al., 2010). Ainsi même si le contenu en gènes semble bien conservé au sein des eudicotylédones et des monocotylédones, ceux-ci ont vraisemblablement subi d'importants réarrangements au cours de leur évolution.

Des études de paléogénomique ont permis d'estimer le nombre de gènes minimal présents dans le génome ancêtre, aussi appelés protogènes. Chez les eudicotylédones, une estimation de 10547 protogènes a été réalisée à partir des génomes de la vigne, du peuplier et d'*Arabidopsis* (Jaillon et al., 2007). Chez les monocotylédones, une estimation de 9138 protogènes a été réalisée à partir des génomes du riz, du sorgho, du maïs, de l'orge et du blé (Salse et al., 2009). Pour l'ancêtre de angiospermes, une première étude estime que 8121 gènes sont conservés entre la vigne, le peuplier, *Arabidopsis* et le riz (Jaillon et al., 2007) et une seconde a établi que les gènes communs entre le sorgho, le riz, *Arabidopsis* et le peuplier sont répartis dans 9503 familles (Paterson et al., 2009). Récemment, l'étude approfondie de ces gènes conservés entre les différents génomes d'angiosperme, a permis de définir des régions ancestrales potentielles entre une eudicotylédone, la vigne, et une monocotylédone, le riz, malgré d'importants réarrangements (Tang et al., 2010).

En conclusion, les caractéristiques de l'espace génique, à savoir le nombre de gènes, leur taille et leur nature, sont relativement constantes entre les génomes de plantes dont la taille est extrêmement variable. De plus, l'espace génique a subi d'importants réarrangements structuraux au cours de l'évolution des génomes végétaux. Ainsi il est probable que les gènes ne soient pas répartis selon le même modèle dans le génome de ces espèces.

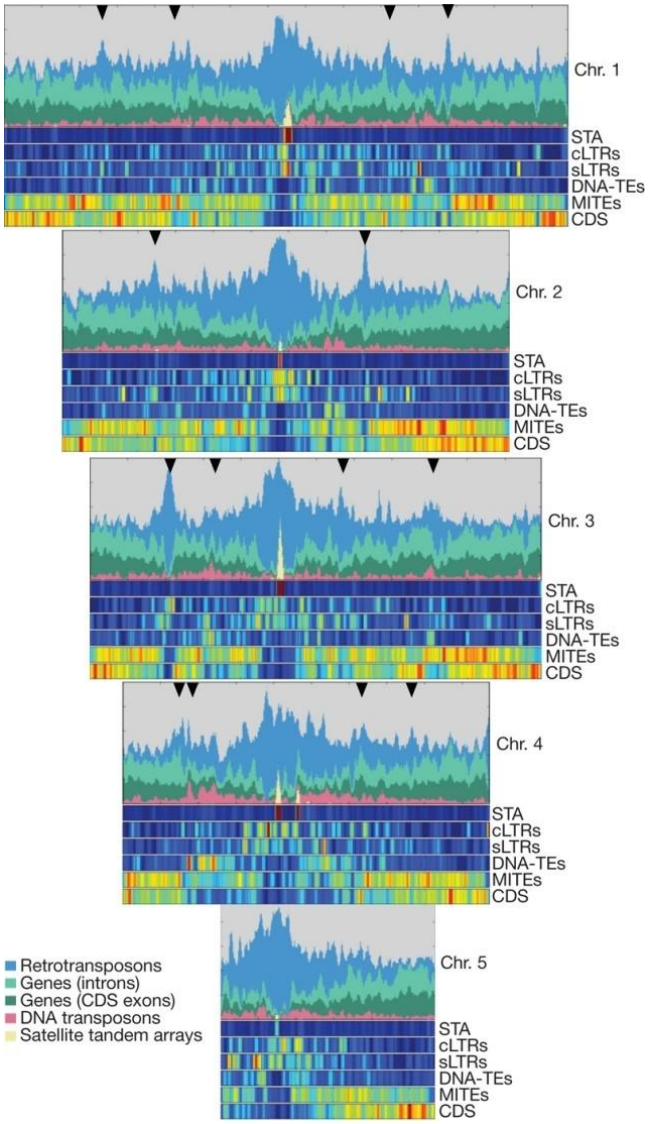


Figure 13 : La distribution chromosomique des principaux composants du génome de *Brachypodium*. L'abondance et la distribution des éléments suivants sont présentés : rétrotransposons à LTR entiers (cLTRs) ; solo-LTR (sLTRs) ; transposons à ADN potentiellement autonomes à l'exception des « Miniature Inverted-repeat Transposable Elements » (MITEs) (DNA-TEs) ; MITEs ; exons des gènes (CDS) ; introns des gènes and « satellite tandem arrays » (STA). Les graphiques représentent les pourcentages de paires de base occupés par les différents éléments dans des fenêtres glissantes. Les « Heat Map » représentent les variations des pourcentages de paires de base entre la valeur minimale et la valeur maximale pour les différents éléments. Les triangles représentent les ruptures de synténie.
(d'après The International *Brachypodium* Initiative, 2010)

2.2. La distribution des gènes dans les génomes de plantes

Il est désormais communément accepté que l'ordre des gènes dans les génomes eucaryotes n'est pas aléatoire mais que l'organisation de l'espace génique est fonction de la taille des génomes (Hurst et al., 2004).

2.2.1. Entre les chromosomes

D'après l'analyse des génomes de plantes séquencés, le contenu en gènes serait réparti de façon homogène entre les chromosomes en fonction de leur taille (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010; The International *Brachypodium* Initiative, 2010). Seules quelques exceptions ont été constatées mais concernent un très petit nombre de chromosomes par génome. Par exemple, le bras court du chromosome 5 de *Brachypodium* est deux fois moins dense en gènes mais plus riche en ET que les autres bras de chromosomes (Figure 13) (The International *Brachypodium* Initiative, 2010). Cette exception à propos de la densité de gènes est également trouvée en position orthologue dans les génomes du riz et du sorgho avec les bras courts des chromosomes 4 et 6 respectivement (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; The International *Brachypodium* Initiative, 2010). Un autre exemple concerne le chromosome 1 dans le génome du peuplier qui est significativement moins concentré en euchromatine que les autres chromosomes (Tuskan et al., 2006). Hormis ces exceptions, aucune réelle différence n'est observable entre les génomes végétaux du point de vue de la répartition des gènes entre les chromosomes.

2.2.2. Le long des chromosomes

Dans les génomes eucaryotes, quelle que soit leur taille, une règle générale demeure quant à la répartition des gènes le long des chromosomes. En effet, les gènes sont répartis le long des bras des chromosomes mais leur densité chute très fortement à proximité des centromères fonctionnels (Lomiento et al., 2008).

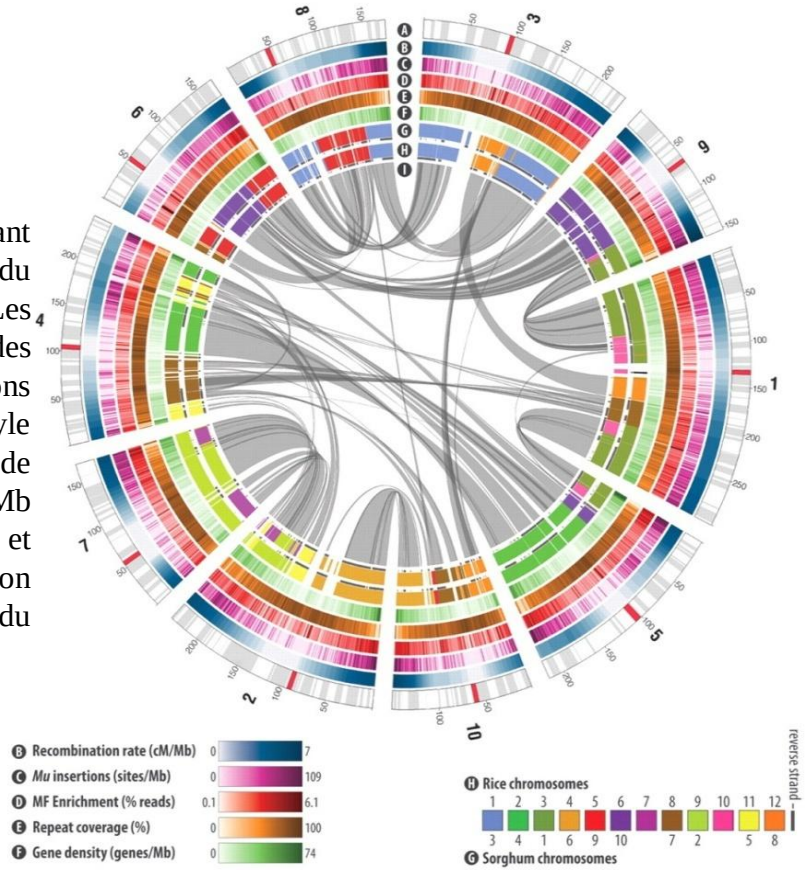


Figure 14 : Les cercles concentriques décrivant les caractéristiques du génome de référence du maïs B73. La structure chromosomique (A). Les positions centromériques sont indiquées par des bandes rouges. Carte génétique (B). Insertions d'ET *Mu* (C). Les lectures de filtration méthyle (D). Séquences répétées (E). Gènes (F). Densité de gènes dans des fenêtres glissantes de 1 Mb espacées de 200 kb. Synténie avec le sorgho (G) et le riz (H). Carte d'homéologie (I). Identification des blocs de gènes dupliqués dans le génome du maïs. (d'après Schnable et al., 2009)

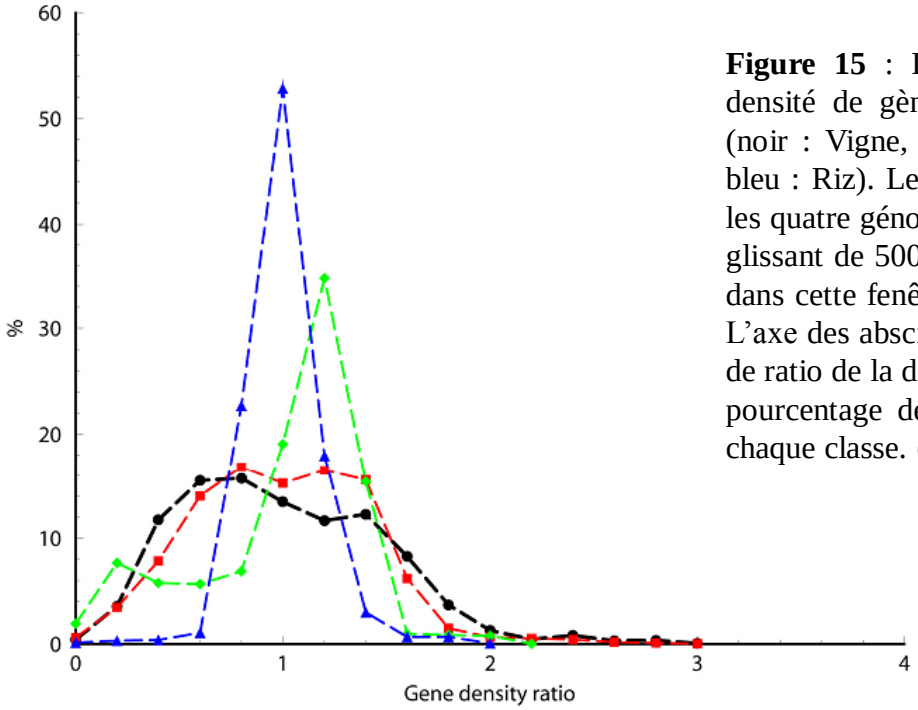


Figure 15 : Distribution de l'homogénéité de la densité de gènes dans quatre génomes de plantes (noir : Vigne, rouge : Peuplier, vert : *Arabidopsis*, bleu : Riz). Le ratio de la densité de gènes est dans les quatre génomes comme suit. Pour chaque fenêtre glissant de 500 kb, le ratio entre la densité de gènes dans cette fenêtre et la densité moyenne est calculé. L'axe des abscisses représente les différentes classes de ratio de la densité de gènes l'axe des ordonnées le pourcentage de fenêtres glissantes correspondant à chaque classe. (d'après Jaillon et al., 2007)

Cependant des répartitions différentielles des gènes sont généralement observées le long des bras de chromosomes en fonction de la taille des génomes. Ainsi, dans les génomes de petite taille comme pour *Chlamydomonas* (121 Mb), *Arabidopsis* (125 Mb), *Brachypodium distachyon* (272 Mb) et le riz (389 Mb), les gènes sont répartis de façon uniforme et homogène le long de leurs chromosomes (Figure 13) (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Merchant et al., 2007; The International *Brachypodium* Initiative, 2010). Pour les génomes de taille intermédiaire comme pour le peuplier (485 Mb) et la vigne (487 Mb), des alternances entre des régions riches en gènes et des régions pauvres en gènes ont pu être observées (Tuskan et al., 2006; Jaillon et al., 2007). Dans le génome du sorgho (730 Mb) et de la pomme (740 Mb), les régions distales des chromosomes semblent plus riches en gènes que les régions proximales (Paterson et al., 2009; Velasco et al., 2010). Cette tendance est encore plus prononcée dans les génomes du soja (1115 Mb) et du maïs (2300 Mb) dans lesquels les gènes sont organisés selon un gradient positif de la densité de gènes des centromères vers les télomères des chromosomes (Figure 14) (Schnable et al., 2009; Schmutz et al., 2010).

De plus, à une échelle plus fine, les gènes ne sont pas répartis de la même façon selon les génomes. D'après la figure 15, les distances intergéniques dans les génomes d'*Arabidopsis* et du riz semblent constantes alors que les distances intergéniques dans les génomes du peuplier et de la vigne suivent une distribution bimodale (Jaillon et al., 2007). Ainsi, certains gènes seraient plus proches les uns des autres que la moyenne, c'est-à-dire organisés en îlots de gènes, et les autres seraient plus éloignés que la moyenne, c'est-à-dire plus isolés.

Malgré l'accumulation de séquences de génomes complets de plantes, les études les plus approfondies concernant l'organisation de l'espace génique à l'échelle de la séquence ont été réalisées à partir de clones BAC individuels ou de régions plus larges composées de BAC chevauchants, aussi appelées contigs. Par exemple, Clough et al. (2004) ont identifié 11 gènes soit un gène tous les 9,4 kb en moyenne dans la séquence d'un BAC de soja de 103 kb. Cette densité correspond à une région riche en gènes chez le soja ce qui suggère une répartition inégale des gènes le long des chromosomes de soja (Clough et al., 2004). Guo et al. (2008) ont analysé la séquence de 142 BAC individuels de coton d'une taille moyenne de 100 kb dont 21% ne portaient pas de gènes mais des ET insérés les uns dans les autres. La densité moyenne de gènes était de un gène tous les 34,5 kb mais avec une fluctuation importante selon les BAC entre 0 et 33,2 gènes par 100 kb. En définissant les îlots de gènes comme des

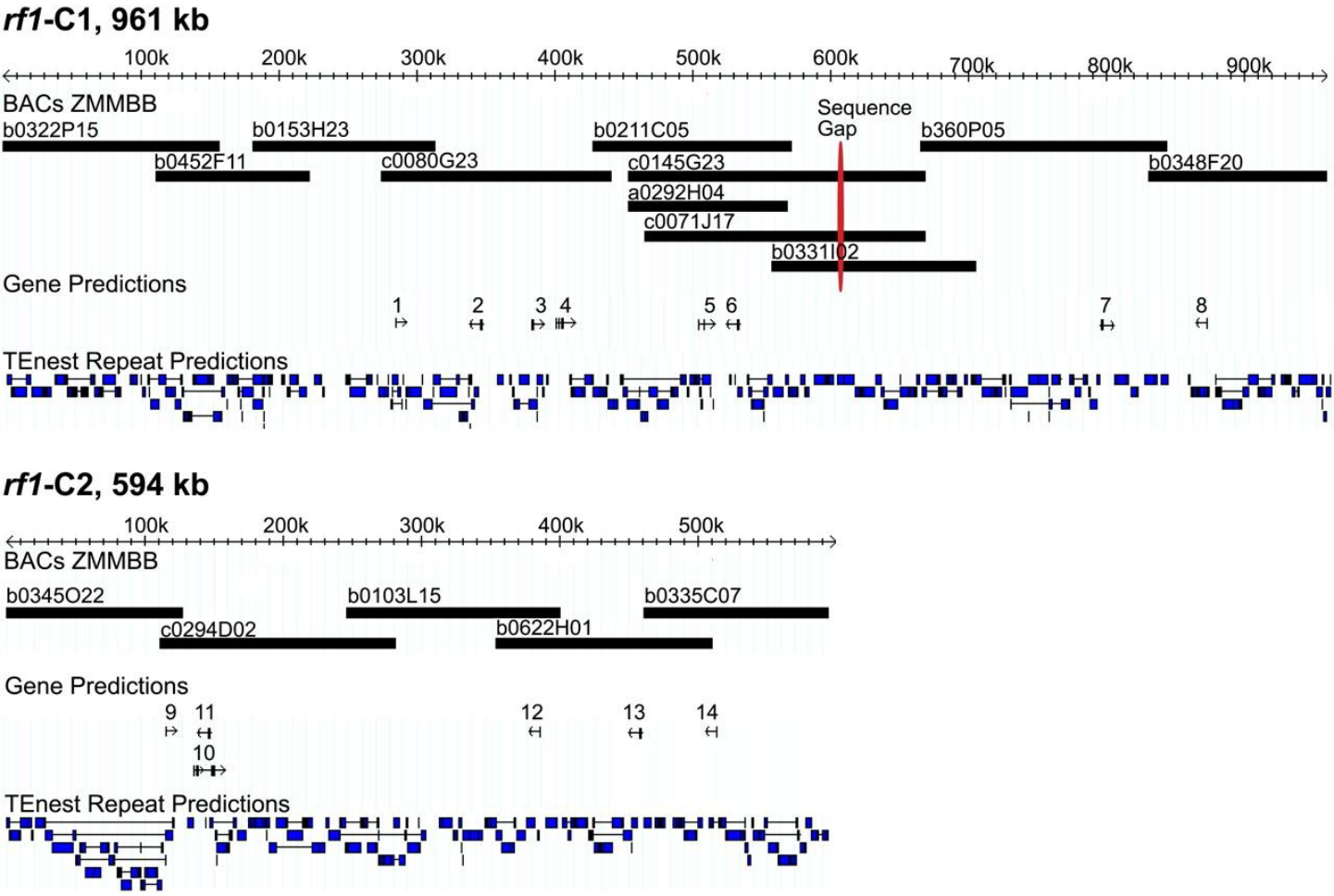


Figure 16 : Carte physique et annotation des séquences de contigs de maïs. Cet affichage du contig *rf1* porté par le chromosome 3 du maïs présente l'ordonnancement des BAC, les prédictions de gènes et l'annotation des ET. *rf1*-C1 a une taille de 961 kb et contient 11 clusters d'ET et huit prédictions de gènes. *rf1*-C2 a une taille de 594 kb et contient cinq clusters d'ET et six prédictions de gènes. (d'après Kronmiller et Wise, 2009)

groupes d'au moins deux gènes séparés par moins de 5kb, Guo et al. (2008) ont mis en évidence qu'au moins 49% des gènes étaient isolés dans le génome du coton. Ainsi, à partir de séquences de BAC individuels, des alternances entre gènes en îlots et gènes isolés séparés par des « océans » d'ET ont également été observées dans des génomes de taille importante comme le soja et le coton.

Ces observations ont été facilitées par le séquençage de contigs de BAC. Avec le séquençage de six BAC d'orge chevauchants formant un contig de 439 kb, Wicker et al. (2005) ont identifié, en extrémité du contig, un îlot contenant deux gènes séparés par environ 1 kb. Le reste du contig, soit plus de 300 kb, était constitué presque exclusivement d'ET insérés les uns dans les autres. Chez le maïs, deux contigs de BAC de 961 kb et 594 kb ont été analysés et des îlots de gènes comptant entre un et quatre gènes ont été identifiés (Figure 16) (Kronmiller and Wise, 2008). La densité en gènes des îlots était en moyenne de un gène tous les 16 kb alors que la densité globale des deux contigs était de un gène tous les 111 kb. Les îlots de gènes était séparés par des clusters d'ET insérés les uns dans les autres sur des tailles variant entre 23 et 155 kb. Enfin l'analyse d'un contig de maïs de 22 Mb a révélé une densité de un gène par 40 kb soit une densité plus importante que la moyenne attendue de un gène par 55 kb (Wei et al., 2009). De plus, une distribution bimodale des distances intergéniques a été constatée étant donné que 45% des distances intergéniques comptaient moins de 10 kb et 44% comptaient plus de 20 kb. Certaines distances intergéniques pouvaient être très larges vu que 8% comptaient plus de 100 kb avec une distance maximale de 530 kb. Ces séquences intergéniques étaient également composées d'ET insérés les uns dans les autres (Wei et al., 2009). Ainsi, ces analyses menées sur des séquences de grande taille confirment également que, dans les génomes les plus grands, les gènes ne sont pas répartis de façon uniforme mais sont soit organisés en îlots de gènes soit isolés au milieu d'océans d'ET.

2.3. Les mécanismes impliqués dans la formation des îlots de gènes dans les génomes de grande taille

La densité de gènes le long des chromosomes est souvent corrélée négativement avec la densité d'ET. De plus, les différentes classes et superfamilles d'ET ont des distributions spécifiques le long des chromosomes, certaines étant plus souvent associées aux gènes et

d'autres plus souvent associées à l'hétérochromatine (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Jaillon et al., 2007; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010; The International *Brachypodium* Initiative, 2010). Du fait de ces observations, de nombreuses études ont analysé l'impact des ET sur la distribution des gènes le long des chromosomes et sur la formation des îlots de gènes.

2.3.1. L'impact de la dynamique des ET dans la structuration des génomes végétaux

2.3.1.1. L'insertion uniforme mais la délétion ciblée

De nombreuses hypothèses ont été énoncées à propos de la dynamique des ET sur la structuration des génomes végétaux et plus particulièrement sur la formation d'îlots de gènes dans les régions distales des chromosomes des génomes de grande taille. La première hypothèse repose essentiellement sur les corrélations négatives entre les distributions de la densité d'ET et du taux de recombinaison et entre les distributions de la densité d'ET et de la densité de gènes observées dans certains génomes (Bowers et al., 2005). Ainsi, la recombinaison aurait un impact négatif sur la présence d'ET dans les régions riches en gènes via les mécanismes de recombinaison illégitime et de recombinaison homologue inégale.

La recombinaison a un impact important sur les ET dans certains génomes. En effet, dans le génome de *Brachypodium* au moins 17 Mb ont été perdus via la formation de solo-LTR (The International *Brachypodium* Initiative, 2010). Dans le génome du soja, un nombre de solo-LTR plus important que de rétrotransposons à LTR intacts a été annoté (Schmutz et al., 2010). Dans le génome du riz, une corrélation positive a été identifiée entre la distribution du taux de recombinaison et le nombre de solo-LTR mais aussi avec la densité de gènes, suggérant une suppression préférentielle des rétrotransposons à LTR dans les régions géniques (Tian et al., 2009).

Des études plus approfondies ont été menées sur les ET dans le génome du maïs. A partir de BAC séquencés, Liu et al. (2007) ont identifié deux pics d'amplification des rétrotransposons à LTR dans les régions pauvres en gènes mais un seul dans les régions riches en gènes. Ils ont

ainsi suggéré que cette différence s'expliquait par une suppression des rétrotransposons à LTR plus efficace au niveau des régions riches en gènes. De plus, dans les îlots de gènes identifiés par Kronmiller et Wise (2009) chez le maïs, la moitié des ET était sous forme tronquée.

En outre, l'analyse de la séquence du génome du sorgho a monté que les insertions les plus récentes de rétrotransposons à LTR étaient distribuées de façon aléatoire le long des chromosomes (Paterson et al., 2009). Par contre, les insertions anciennes étaient principalement localisées au niveau des centromères où les gènes sont présents en faible nombre. Ainsi, les auteurs utilisent cette hypothèse de l'insertion aléatoire des rétrotransposons à LTR le long de chromosomes couplée à une délétion ciblée via les mécanismes de recombinaison au niveau des régions riches en gènes pour expliquer leurs observations.

Cependant une étude menée sur *Arabidopsis* a démontré que toutes les superfamilles de rétrotransposons à LTR ne se comportait pas de la même façon face à leur insertion et à leur élimination via les mécanismes de recombinaison (Pereira, 2004). En effet, la dynamique des membres de la superfamille des *copia* répond à l'hypothèse de l'insertion aléatoire et de l'élimination ciblée. Par contre les superfamilles *athila* et *gypsy* auraient un mécanisme de transposition plus ciblée privilégiant leur insertion dans les régions péricentromériques comme détaillée dans le paragraphe suivant (Pereira, 2004).

2.3.1.2. L'ouverture d'océans d'éléments transposables ou l'insertion ciblée

L'hypothèse précédente de l'insertion aléatoire suivie d'une délétion ciblée des rétrotransposons à LTR n'est cependant pas exclusive. D'une part, elle ne s'applique pas nécessairement à toutes les superfamilles d'ET (Pereira, 2004). D'autre part, après le séquençage du génome entier du maïs, elle ne permettait pas d'expliquer les accumulations différentielles de rétrotransposons à LTR observées (Baucom et al., 2009a). En effet, les accumulations de rétrotransposons à LTR imbriqués les uns dans les autres sont fréquentes dans les génomes de grande taille et sont parfois qualifiées d'océans d'ET à opposer aux îlots de gènes (SanMiguel et al., 1998; Kronmiller and Wise, 2009; Schmutz et al., 2010). Les

océans identifiés par Kronmiller et Wise (2009) s'étendaient de 23 à 155 kb et étaient composés de 3 à 18 insertions. Les imbrications se font majoritairement entre rétrotransposons à LTR de familles différentes (Liu et al., 2007).

Plus généralement, la seconde hypothèse quant à l'impact de la dynamique des ET sur la structuration de l'espace génique et de la formation d'îlots de gènes, serait que les ET s'insèrent de façon préférentielle et spécifique (Baucom et al., 2009a). Par exemple, l'élément *Basho*, identifié dans le génome d'*Arabidopsis*, s'insèrerait préférentiellement au niveau de sites poly(A) (The *Arabidopsis* Genome Initiative, 2000). D'autres ET comme les rétrotransposons à LTR des superfamilles *athila* et *gypsy* cibleraient préférentiellement les régions hétérochromatiques des chromosomes (Pereira, 2004). La distribution du transposon à ADN *Mutator* chez le maïs est corrélée avec la densité de gènes et le taux de recombinaison (Liu et al., 2009). Cet élément *Mutator* privilégierait les régions à structure chromatinienne ouverte comme site d'insertion donc la proximité des gènes actifs (Liu et al., 2009).

La spécificité d'insertion des ET serait conférée par leur intégrase. En effet, les rétrotransposons à LTR du type *copia* possèdent une intégrase dont la partie C terminale a des motifs différents des rétrotransposons à LTR du type *athila* ou *gypsy* (Pereira, 2004). Chez la levure, en effet, Gao et al. (2008) ont démontré par la technique de « protéine fusion » que la partie C terminale de l'intégrase permettait à un rétrotransposon de type *gypsy* de cibler les modifications d'histones associées à un état hétérochromatique de la région. En outre, la spécificité de cette partie C terminale de l'intégrase semble primordiale pour la reconnaissance du site d'insertion puisque la modification d'un unique acide aminé à ce niveau de l'intégrase d'un rétrotransposon chez la levure modifie complètement son profil d'insertion dans le génome (Gai and Voytas, 1998).

En conclusion, les ET sont des facteurs importants de structuration des génomes et de l'espace génique en particulier. Ils agissent soit de façon directe, en s'insérant préférentiellement dans des régions spécifiques du génome, ou de façon indirecte, en étant éliminer de façon préférentielle des régions géniques à fort taux de recombinaison (Figure 18A et B). Les ET peuvent également agir de façon plus directe sur la répartition des gènes en provoquant des réarrangements et des duplications de gènes.

2.3.2. Les duplications de gènes

2.3.2.1. Le « transport » de gènes via les éléments transposables

Un autre mécanisme susceptible d'intervenir dans la formation d'îlot de gènes est le « transport » de gènes. De nombreux réarrangements de gènes ont été identifiés par l'étude de la synténie entre les génomes comme présenté précédemment. Les ET sont souvent considérés comme les principaux responsables des réarrangements des gènes (Kapitonov and Jurka, 2007). En effet, dans de nombreux génomes de plantes, des ET semblent avoir capturé des gènes ou des fragments de gènes (The *Arabidopsis* Genome Initiative, 2000; Baucom et al., 2009a; Schnable et al., 2009). Des fragments de gènes ont été observés dans des rétrotransposons à LTR, des CACTA, des *helitrons* et dans des *Mutators* appelés Pack-MULE (Dooner and Weil, 2007; Baucom et al., 2009a; Paterson et al., 2009). De plus, dans le génome du maïs, les *helitrons* et les *Mutators*, les ET les plus efficaces dans la capture de gènes, ont une distribution biaisée en faveur des régions riches en gènes (Schnable et al., 2009). Cette répartition suggère ainsi que la capture de gènes est plus efficace au niveau des régions riches en gènes. Cependant il semble que des disparités existent entre les génomes quant à l'efficacité de la capture de gènes par les ET. Par exemple, Sweredoski et al. (2008) ont trouvé que les *helitrons* capturaient plus de fragments de gènes dans le génome du maïs que dans le génome du riz.

Certaines études ont montré que les ET ne favorisaient pas seulement la duplication de gènes mais également l'expression de ces gènes capturés (Akhunov et al., 2007; Hanada et al., 2009). Dans le génome du riz, 22% des Pack-MULE étudiés contenant des gènes, étaient transcrits et 28 d'entre eux étaient même traduits (Hanada et al., 2009). De plus, les Pack-MULE portant des fragments issus de différents gènes, étaient plus fréquemment exprimés que les autres ne contenant que des fragments issus du même gène. Enfin, la répression de ces éléments a été étudiée et lorsqu'un Pack-MULE générait un siRNA, l'expression de l'élément et du gène initial était réduite (Hanada et al., 2009). Ainsi, la capture de gènes par les ET est de plus en plus considérée comme une stratégie de contournement des mécanismes de répression mis en place par l'organisme hôte contre les ET (Lisch, 2009; Tenaillon et al., 2010).

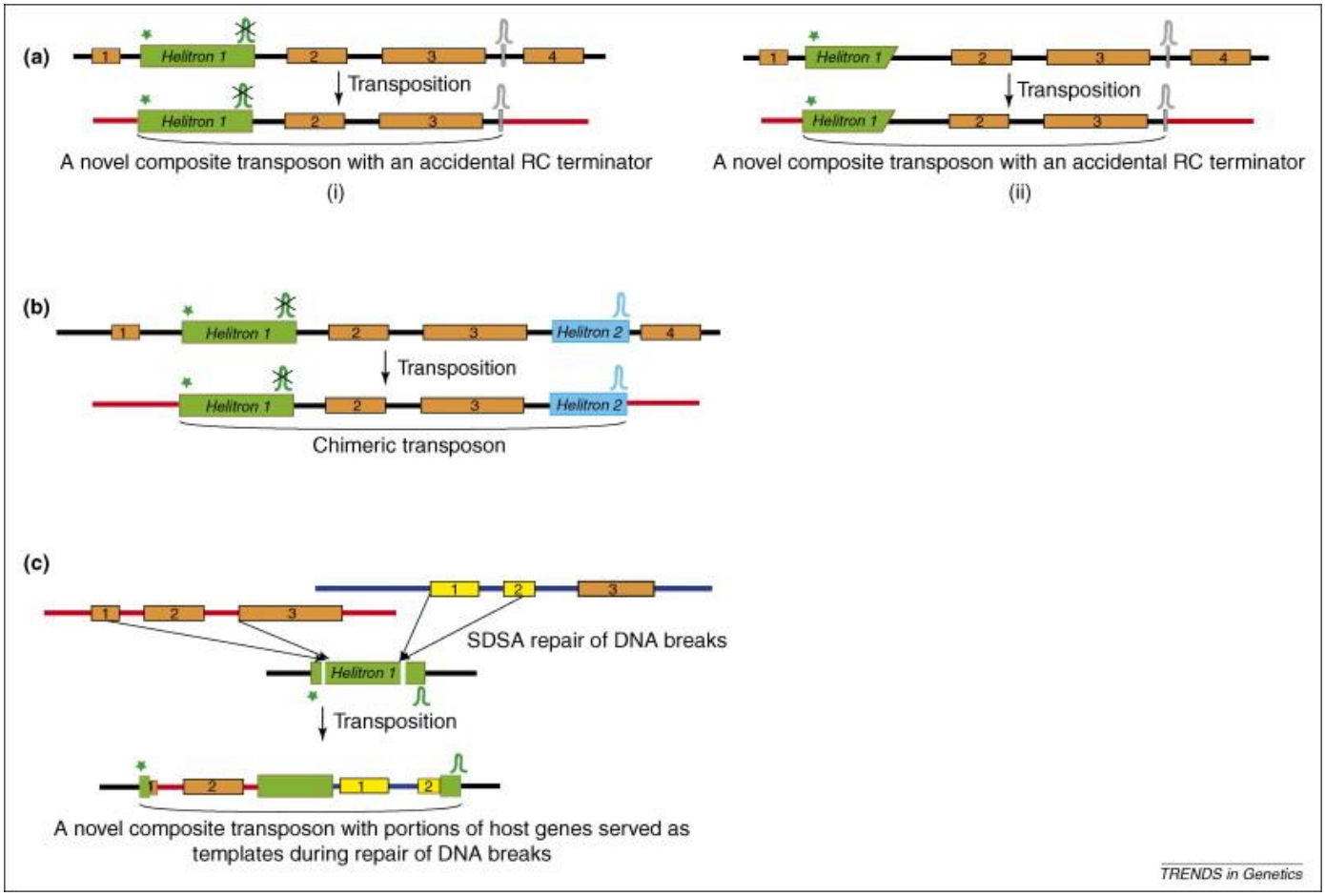


Figure 17 : Différents modèles de captures de gène par les *Helitrons*. (a) Dans le premier modèle « read- through », le fragment exon2-intron-exon3 du gène peut être copié par un transposon à cause (i) d'une anomalie de la séquence terminatrice de l'*Helitron 1* ou (ii) d'une troncature de l'extrémité 3' de l'*Helitron*. La partie terminatrice défectueuse peut être remplacée par une séquence similaire dans l'intron après l'exon 3. (b) Dans le second modèle « read- through », le fragment exon2-intron-exon3 du gène est copié entre les *Helitron* 1 et 2 car la séquence terminatrice de l'*Helitron* 2 a été utilisée. (c) Dans le modèle « filler DNA », deux gènes issus de différents chromosomes sont utilisés pour réparer les cassures double-brin qui ont eu lieu dans l'*Helitron* 1. Dans tous les modèles, les astérisques indiquent l'extrémité 5' fonctionnelle des Helitrons alors que les boucles et les boucles barrées représentent respectivement les séquences terminatrices fonctionnelles et non fonctionnelles. (d'après Kapitonov et Jurka, 2007)

Comme rapporté par Kapitonov et Jurka (2007), l'hypothèse avait été émise que la capture de gènes serait réalisée au moment de la transposition et résulterait d'une excision imparfaite. En effet, la coupure serait réalisée plus loin que l'extrémité de l'ET et celui-ci transposerait avec un fragment d'ADN supplémentaire pouvant contenir des gènes ou des fragments de gènes (Figure 17a et b) (Kapitonov and Jurka, 2007). Cependant de récentes études suggèrent que les ET ne seraient pas les responsables directs de ces captures de fragments de gènes ou de gènes entiers (Dooner and Weil, 2007; Kapitonov and Jurka, 2007; Wicker et al., 2010). En effet, les ET seraient les initiateurs de cassures double-brin lors de la transposition et ces dernières seraient réparées par les mécanismes classiques de recombinaison homologue ou de recombinaison illégitime (Figure 17c) (Puchta, 2005; Dooner and Weil, 2007; Kapitonov and Jurka, 2007; Kass and Jasin, 2010; Wicker et al., 2010). Aussi, lors de réparations imparfaites, le fragment d'ADN utilisé pour combler la cassure peut porter un gène ou un fragment de gène (Wicker et al., 2010).

2.3.2.2. Les gènes dupliqués en tandem

Des gènes dupliqués en tandem, dont les copies ont été maintenues côte à côte ou proche, sont observés dans tous les génomes de plantes séquencés et représentent entre 10 et 20% du contenu en gènes en fonction des espèces et des critères de définition (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). La plupart des groupes de gènes dupliqués en tandem compte deux gènes mais les plus grands groupes peuvent compter plusieurs dizaines de gènes (The *Arabidopsis* Genome Initiative, 2000; Zhang and Gaut, 2003; Rizzon et al., 2006; Tuskan et al., 2006). La proportion de gènes dupliqués en tandem par rapport aux gènes présents en simple copie est significativement plus faible au niveau des centromères dans les génomes d'*Arabidopsis* et du riz (Zhang and Gaut, 2003; Rizzon et al., 2006). Dans ces mêmes génomes, le nombre de gènes dupliqués en tandem est également corrélé avec le taux de recombinaison (Zhang and Gaut, 2003; Rizzon et al., 2006). De plus, chez les mammifères, ces gènes tendent à former des îlots de gènes puisqu'ils sont plus proches les uns des autres que les gènes sans homologie (Shoja and Zhang, 2006).

Les gènes dupliqués en tandem sont généralement produits lors de réparations de cassures double-brin grâce à des mécanismes similaires à ceux mis en place lors de cassures double-brin liées aux ET (Helleday, 2003; Fiston-Lavier et al., 2007). Ainsi, les réparations de cassures double-brin, au même titre que les ET, ont des impacts importants sur la structuration de l'espace génique puisqu'elles génèrent des duplications et des réarrangements de gènes surtout en dehors de régions centromériques (Figure 18C et D).

2.3.3. Des gènes conservés proches pour des raisons fonctionnelles

Cependant outre les aspects et mécanismes structuraux, de nombreux gènes sont organisés en îlots car ils possèdent des caractéristiques fonctionnelles communes entre gènes proches. Par exemple, dans les génomes du riz et d'*Arabidopsis*, entre 2 et 10% des gènes adjacents sont coexprimés par groupe de deux à quatre gènes (Ren et al., 2005; Zhan et al., 2006; Ren et al., 2007). Zhan et al. (2006) précisent que les gènes coexprimés sont nécessairement proches mais pas forcément strictement adjacents. Une fraction importante de ces gènes proches coexprimés est issue de duplications en tandem (Williams and Bowles, 2004; Schmid et al., 2005; Zhan et al., 2006). Ainsi, les gènes dupliqués en tandem auraient deux fois plus de chance d'être coexprimés que les gènes adjacents sans homologie entre eux (Ren et al., 2005). Cependant une partie significative de gènes proches coexprimés ne partagent pas de relation d'homologie (Ren et al., 2005; Zhan et al., 2006).

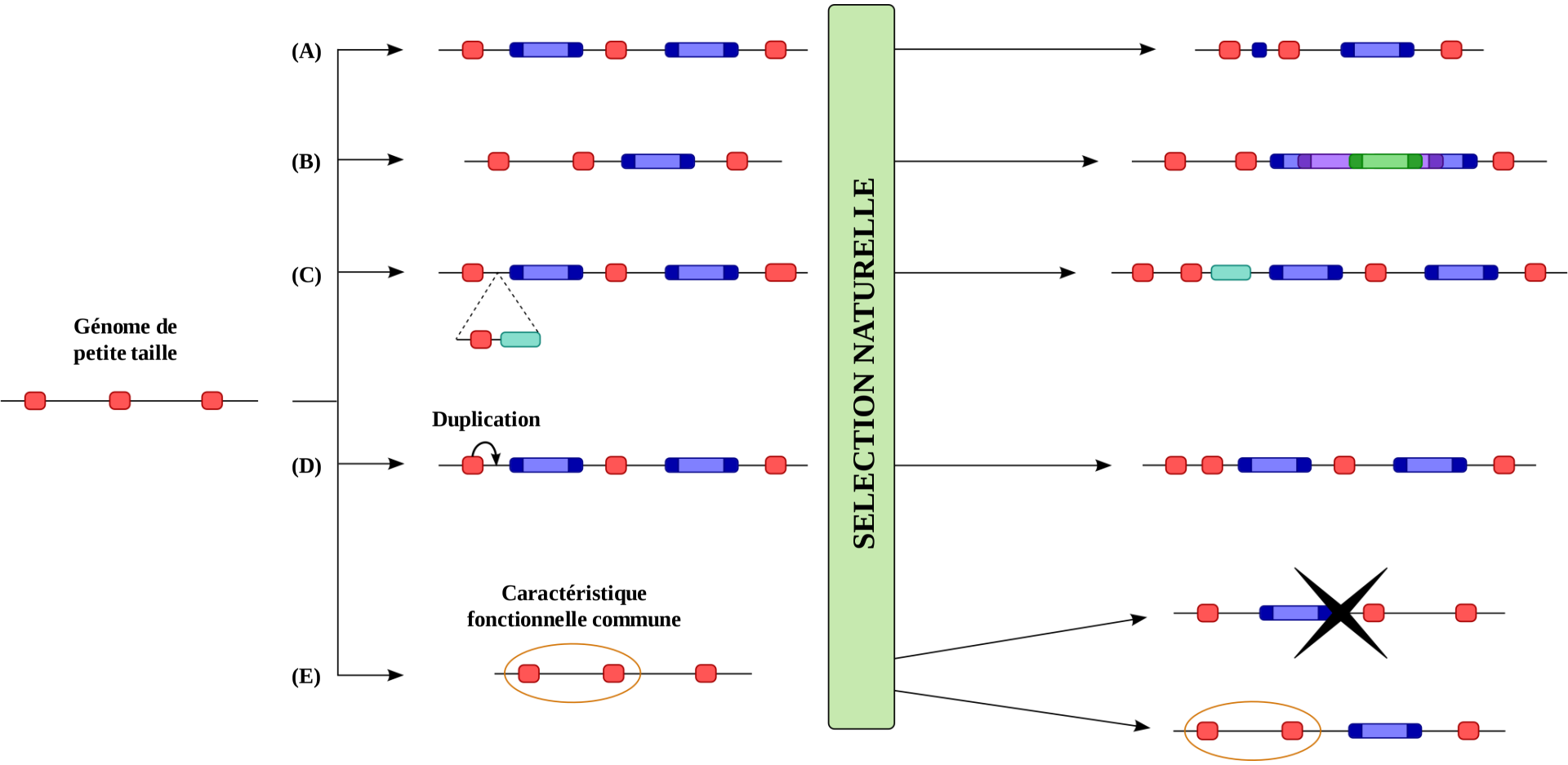
En plus d'être coexprimés, certains gènes en îlots partagent la même fonction ou sont impliqués dans le même processus biologique (Williams and Bowles, 2004; Schmid et al., 2005; Zhan et al., 2006; Liu and Han, 2009). Les gènes coexprimés semblent même biaisés en faveur de certaines fonctions particulières comme les protéines du métabolisme ou les protéines impliquées dans des complexes du type protéasome ou ribosome (Williams and Bowles, 2004; Ren et al., 2005; Schmid et al., 2005). Les gènes organisés en îlots ayant des caractéristiques fonctionnelles communes, peuvent également avoir la même fonction ou faire partie du même processus biologique sans nécessairement être coexprimés comme dans le génome d'*Arabidopsis* ou du coton (Lee and Sonnhammer, 2003; Xu et al., 2008). De plus, des îlots de gènes de ménage ont été identifiés dans le génome humain (Lercher et al., 2002).

De nombreuses études ont observé que les gènes coexprimés présentaient des distances intergéniques significativement plus courtes que les autres paires de gènes ce qui suggèrent un mécanisme de régulation commun (Cohen et al., 2000; Williams and Bowles, 2004; Zhan et al., 2006). En effet, des séquences régulatrices ou des promoteurs communs ont été identifiés comme agissant à courte distance (Cohen et al., 2000; Hurst et al., 2004; Williams and Bowles, 2004; Babu et al., 2008; Chen et al., 2010a). Dans des cas extrêmes de régulations communes, comme identifiés chez *Arabidopsis*, la coexpression de gènes implique parfois des transcrits polycistroniques, c'est-à-dire des ARNm contenant plusieurs cadres ouverts de lecture et donc codant plusieurs protéines (Liu and Han, 2009; Chen et al., 2010a). Les régulations identifiées à plus longues distances impliquent la structure chromatinienne régulée via les modifications des histones (Hurst et al., 2004; Ren et al., 2005; Lercher and Hurst, 2006; Batada et al., 2007; Chen et al., 2010a). Ainsi, les gènes proches soumis aux mêmes variations de structure chromatinienne tendent à être coexprimés (Batada et al., 2007; Gierman et al., 2007). De plus, l'étude menée par Gierman et al. (2007) chez l'humain souligne l'impact majeur de la structure chromatinienne sur la coexpression des gènes puisqu'un transgène inséré à 90 positions différentes dans le génome mimait de façon significative le niveau d'expression des gènes présents au niveau de la région d'insertion.

Ainsi, des îlots de gènes partageant des caractéristiques fonctionnelles communes sont identifiés plus fréquemment que par le seul fait du hasard (Hurst et al., 2004). Ils auraient donc une signification fonctionnelle et la sélection naturelle maintiendrait ces gènes proches les uns des autres car cette configuration apporterait un avantage sélectif aux individus (Figure 18E) (Sémon and Duret, 2006; Batada et al., 2007; Babu et al., 2008). Cette hypothèse est confirmée par la mise en évidence de gènes proches coexprimés ou partageant les mêmes fonctions qui sont conservés entre les génomes d'*Arabidopsis*, du riz et du peuplier (Krom and Ramakrishna, 2008; Liu and Han, 2009).

2.3.4. La sélection naturelle comme arbitre des réarrangements structuraux

Comme développé dans les parties précédentes, les insertions et les délétions d'ET, les réparations de cassures double-brin ou les partages d'une caractéristique fonctionnelle entre



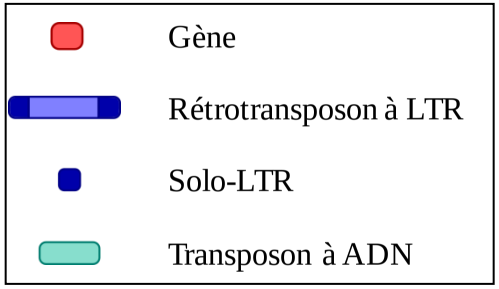


Figure 18 : Les mécanismes de formation des îlots de gènes. (A) L'insertion uniforme mais la délétion ciblée des rétrotransposons à LTR. (B) L'insertion ciblée de rétrotransposons à LTR les uns dans les autres. (C) Le transport de gène par un transposon à ADN et leur insertion à proximité d'un autre gène. (D) La duplication de gène en tandem. (E) Le maintien de paires de gènes proches pour des raisons fonctionnelles. La sélection naturelle s'exerce sur toutes les voies de formation des îlots et peut aussi bien avoir un effet positif, neutre ou négatif sur les nouvelles organisations de l'espace génique.

gènes proches sont autant de facteurs impactant l'organisation de l'espace génique et la formation d'îlots de gènes dans les génomes et plus particulièrement dans les génomes de grande taille (Figure 18).

Cependant, même si ces mécanismes semblent importants dans la structuration des génomes, leurs effets ne sont conservés et transmis que s'ils sont sélectionnés via la sélection naturelle. Ainsi, les modifications structurales liées aux ET ne peuvent être fixées dans une population que si elles sont neutres ou avantageuses du point de vue de la fitness des individus (Kidwell and Lisch, 2001; Wright et al., 2003; Baucom et al., 2009a; Grover and Wendel, 2010; Tenaillon et al., 2010). Ainsi, dans la majorité des cas, les réarrangements, de quelque origine qu'ils soient, seraient délétères et donc contre-sélectionnés dans les régions euchromatiques (Bowers et al., 2005). Par exemple, la plupart des nouvelles insertions d'ET seraient soumises à une forte sélection car elles peuvent avoir des conséquences directes sur la fonctionnalité des gènes si l'ET s'insère dans un exon, un intron ou une séquence intergénique potentiellement régulatrice (Wright et al., 2003; Hollister and Gaut, 2009).

Par contre, l'impact de l'insertion d'ET à proximité ou dans les gènes serait à moduler en fonction de la taille des ET. Ainsi, l'insertion d'ET de grande taille dans les introns altère le fonctionnement du gène alors que des insertions d'ET courts du type MITE dans les introns affecteraient moins les gènes (Oki et al., 2008). Pourtant, malgré leur petite taille, les insertions de MITE dans les exons semblent tout de même contre-sélectionnées par rapport aux insertions dans les introns (Oki et al., 2008).

Enfin, outre les effets délétères d'insertions dans les parties fonctionnelles des gènes, les ET peuvent également être néfastes par leur impact sur l'expression des gènes à leur proximité. En effet, comme énoncé précédemment, la transposition des ET est contenue grâce à la mise en place et au maintien de marques épigénétiques telles que la méthylation de l'ADN ou les modifications d'histones. Cependant, la présence de marques épigénétiques au niveau des séquences d'ET ont un impact sur l'expression des gènes avoisinants (Kashkush and Khasdan, 2007; Hollister and Gaut, 2009). Ainsi, Hollister et Gaut (2009) ont mis en évidence que les gènes proches d'ET méthylés étaient exprimés à un niveau moindre que les gènes proches d'ET non-méthylés. De plus, ils ont constaté que les ET méthylés situés à proximité des gènes étaient contre-sélectionnés alors que les ET non-méthylés ou les ET situés loin des gènes ne l'étaient pas. Enfin, ils ont observé que les vieux ET méthylés sont significativement situés

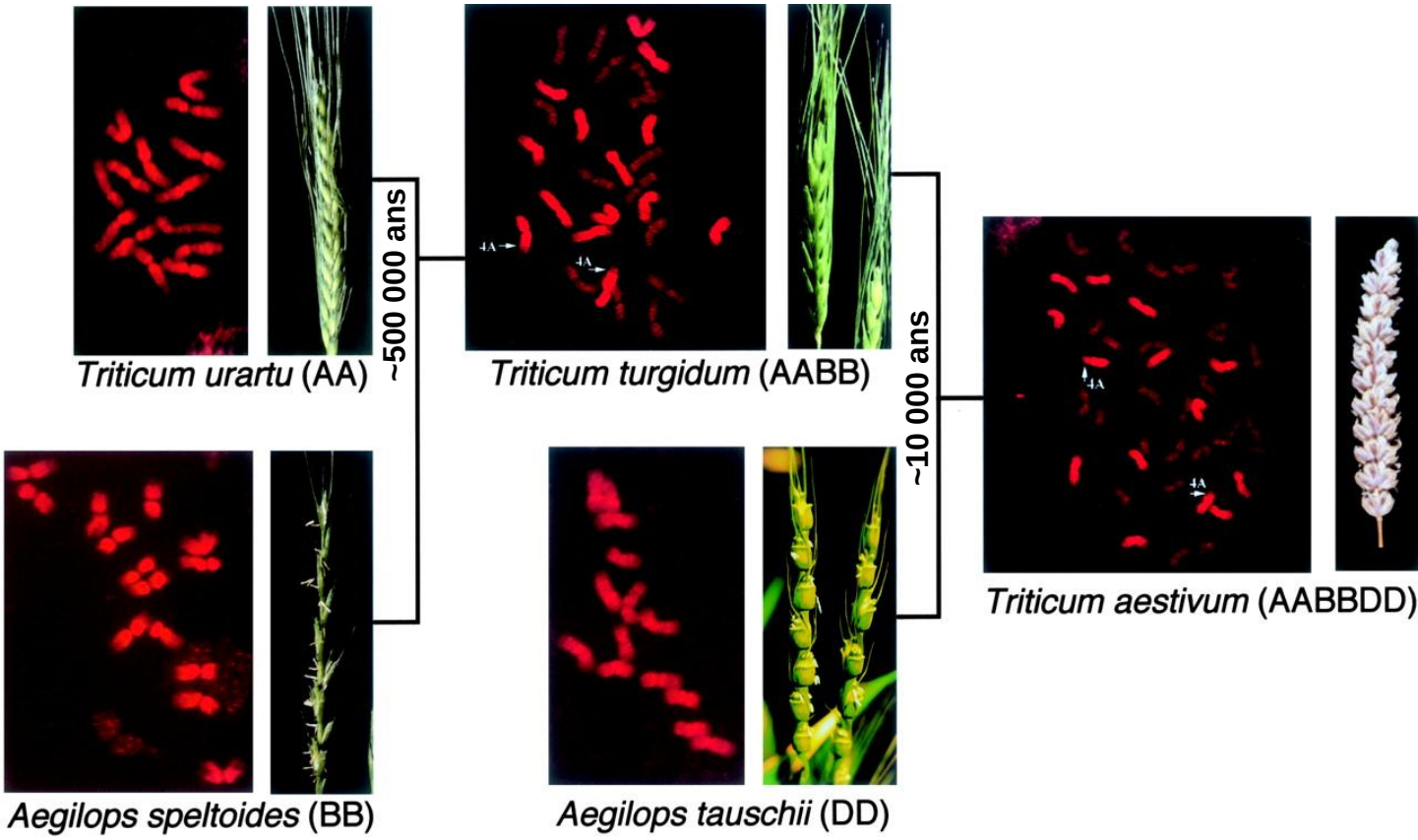


Figure 19 : Les évènements de polyploïdie à l'origine de la formation du blé tendre. (d'après Gill et al., 2004)

loin des gènes. Suite à ces observations, Hollister et Gaut (2009) ont suggéré que la méthylation des ET proches des gènes augmenterait leur effet délétère sur l'expression de ces gènes et que ceci a pour conséquence une perte préférentielle des ET méthylés dans les régions riches en gènes. De plus, les ET courts comme les MITE, les SINE ou les *helitrons*, sont proportionnellement moins méthylés que les ET longs (Hollister and Gaut, 2009). Ainsi les ET de petite taille seraient moins soumis à contre-sélection dans les régions riches en gènes car leur insertion altèrerait moins la fonctionnalité des gènes d'une part et leur expression d'autre part.

En conclusion, en agissant contre les effets délétères des réarrangements structuraux, la sélection naturelle favoriserait le cloisonnement des grands ET tels que les rétrotransposons à LTR dans des régions éloignées des gènes mais aussi la conservation des distances intergéniques et donc des îlots de gènes.

3. L'ORGANISATION DE L'ESPACE GENIQUE CHEZ LE BLE TENDRE

3.1. La structure complexe du génome du blé tendre

Malgré son importance majeure dans l'agriculture et l'alimentation mondiale, la recherche et la sélection sur le blé tendre sont considérablement ralenties par la complexité de son génome, notamment sa taille estimée à 17,5 pg soit 17,1 Gb (Zonneveld et al., 2005).

Cette grande taille s'explique en partie par le fait que le blé tendre est un allohexaploïde récent qui compte trois sous-génomes homéologues, A, B et D (2n = 6x = 42 chromosomes). Il y a environ 3 millions d'années, les ancêtres des génomes A, B et D divergeaient d'un ancêtre commun. Puis, il y a environ 500000 ans, l'hybridation interspécifique entre *Triticum urartu* (génome A) et une espèce proche d'*Aegilops speltoides* (génome S, potentiel ancêtre du génome B) forme le tétraploïde *Triticum turgidum* spp. *dicoccoides* (génome AB), forme sauvage du blé dur (Figure 19) (Gill et al., 2004; Dubcovsky and Dvorak, 2007). L'origine du génome B reste controversée et de nombreuses études ont déjà été menée pour approfondir notre connaissance de son ancêtre (Petersen et al., 2006; Salse et al., 2008b). Enfin, il y a environ 10000 ans, la forme cultivée du blé tétraploïde, *Triticum turgidum* spp. *dicoccum*

Tableau 2 : Les estimations du contenu en ET dans le génome du blé et dans le génome du maïs.

	Chromosome 3B (Paux et al., 2006)	**Génome D** (Paux et al., 2006)	**Chromosome 3B** (Charles et al., 2008)	**Chromosome 3B** (Choulet et al., 2010)	**Maïs** (Schnable et al., 2009)
ET classe I	68,7%	48,0%	61,9%	65,9%	75,6%
Rétrotransposons à LTR	67,4%	47,0%	59,5%	64,1%	74,6%
gypsy	38,7%	24,4%	30,8%	43,6%	46,4%
copia	14,0%	13,7%	14,7%	16,7%	23,7%
Rétrotransposons sans LTR	1,3%	1,1%	2,4%	1,8%	1,0%
LINE	1,3%	1,1%	2,2%	1,7%	1,0%
ET classe II	5,6%	11,6%	16,2%	14,5%	8,6%
CACTA	4,9%	10,8%	15,5%	13,6%	3,2%
MITE	0,5%	0,4%	0,5%	0,2%	0,1%
ET autres et inconnus	1,6%	2,8%	1,0%	1,5%	/
TOTAL	**75,9%**	**62,4%**	**79,1%**	**81,9%**	**84,2%**

(génome AB), s'hybride avec *Aegilops tauschii* (génome D) dans la région du Croissant Fertile pour former le blé tendre, *Triticum aestivum* (génome ABD) (Gill et al., 2004; Sabot et al., 2005; Dubcovsky and Dvorak, 2007). Suite à sa formation, le génome du blé aurait subi des réarrangements structuraux et fonctionnels conséquents et soudains aidant à sa stabilisation (Feldman and Levy, 2009). En effet, les polyploïdes naturels de blé auraient des sous-génomes entre 2% et 10% plus petits que leurs progéniteurs diploïdes (Feldman and Levy, 2009). Ces réarrangements seraient reproductibles sachant que l'élimination de séquences identiques a été observée chez des polyploïdes naturels et des synthétiques (Feldman and Levy, 2009). Cependant, une partie conséquente des changements se ferait également au niveau de la régulation des gènes sans perte de séquence (Feldman and Levy, 2009). Ainsi, le génome du blé tendre, en tant qu'hexaploïde récent, serait resté fortement redondant du point de vue de ses gènes permettant la mise en place de régulations fines et de compensation de dose (Qi et al., 2002; Feldman and Levy, 2009).

Outre la polyploïdie, les génomes diploïdes qui composent le génome du blé tendre sont de taille importante par rapport aux autres espèces de céréales cultivées comme le riz (389 Mb), le sorgho (730 Mb) ou le maïs (2300 Mb) (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009). En effet, *Triticum urartu* et *Aegilops speltoides* ont des tailles de génome de respectivement 4800 Mb et 5000 Mb (Plant DNA C-values database). Cette taille importante est corrélée à une forte proportion d'ET dans le génome du blé. En effet, les ET représenteraient entre 62% et 82% de la taille totale du génome du blé selon les études (Paux et al., 2006; Charles et al., 2008; Choulet et al., 2010). Comme dans les grands génomes précédemment cités, les ET de classe I sont largement majoritaires puisqu'ils représentent entre 48% et 69% des séquences annotées de blé. Les rétrotransposons à LTR sont les principaux représentants de cette classe et représentent entre 47% et 67%. Les superfamilles *gypsy* et *copia* sont respectivement les plus abondantes. Les rétrotransposons sans LTR dont les LINE sont les principaux représentants, ne représentent qu'entre 1% et 2% du génome du blé. Les ET de classe II, dont les CACTA représentent la plus grande proportion en taille, composent entre 6% et 16% du génome de blé. Parmi cette classe, les MITE ne représentent qu'environ 0,5% malgré leur nombre de copies important dans le génome du blé (Paux et al., 2006; Charles et al., 2008; Choulet et al., 2010). Ainsi, la composition du génome du blé est comparable à celle du génome du maïs, le plus grand génome séquencé à l'heure actuelle (Tableau 2) (Schnable et al., 2009).

En conclusion, le génome du blé, du fait de sa taille, de son allohexaploïdie et de son contenu en ET, est un des génomes les plus complexes parmi les plantes cultivées à ce jour. Malgré les développements technologiques récents, il représente toujours un défi pour les analyses moléculaires et la séquence du génome complet n'est toujours pas disponible. Même si la composition et la répartition des ET et des gènes dans le génome du blé correspondent à celles d'un grand génome comme celui du maïs, dont la séquence du génome complet a été publiée il y a environ un an, la polyploïdie reste une difficulté majeure.

3.2. L'espace génique chez le blé tendre, un sujet controversé

3.2.1. Les estimations du contenu en gènes

De nombreuses études ont tenté d'estimer le nombre de gènes présents dans le génome du blé. Certains auteurs se sont basés sur les données de génomique comparative chez les céréales pour estimer le nombre de gènes dans le génome du blé. En effet, la conservation des gènes entre les génomes des céréales est bonne et certaines ressources génomiques peuvent être transférables d'une espèce à l'autre (Moore et al., 1995; Keller and Feuillet, 2000; Devos, 2005; Bolot et al., 2009). Ainsi des études ont utilisé le nombre de gènes obtenu dans les séquences de génomes d'espèces modèles comme le riz, *Brachypodium* ou même *Arabidopsis* pour estimer que les sous-génomes du blé devaient compter environ le même nombre de gènes soit entre 25000 et 50000 gènes (Sidhu and Gill, 2004; Devos, 2010).

L'extraction de la fraction hypométhylée du génome du blé a permis à Rabinowicz et al. (2005) d'estimer que le génome du blé comptait plus de 98000 gènes par sous-génome. L'annotation de séquences d'extrémités de BAC a restreint l'estimation du contenu en gènes à environ 38000 par sous-génome (Paux et al., 2006). Cependant les études récentes basées sur l'annotation de séquences de BAC ou de contigs estiment plutôt le contenu en gènes aux environs de 50000 gènes par sous-génome (Devos et al., 2005; Choulet et al., 2010). Ces deux dernières estimations correspondent à la gamme de 38000 à 48000 gènes estimée pour le contenu génique chez l'orge, espèce proche du blé ayant divergé depuis environ 10 millions d'années (Mayer et al., 2009). Ainsi, comme le soja ou le maïs, tous deux d'anciens polyploïdes, le génome du blé compterait un peu plus de gènes que les autres espèces végétales à petits génomes.

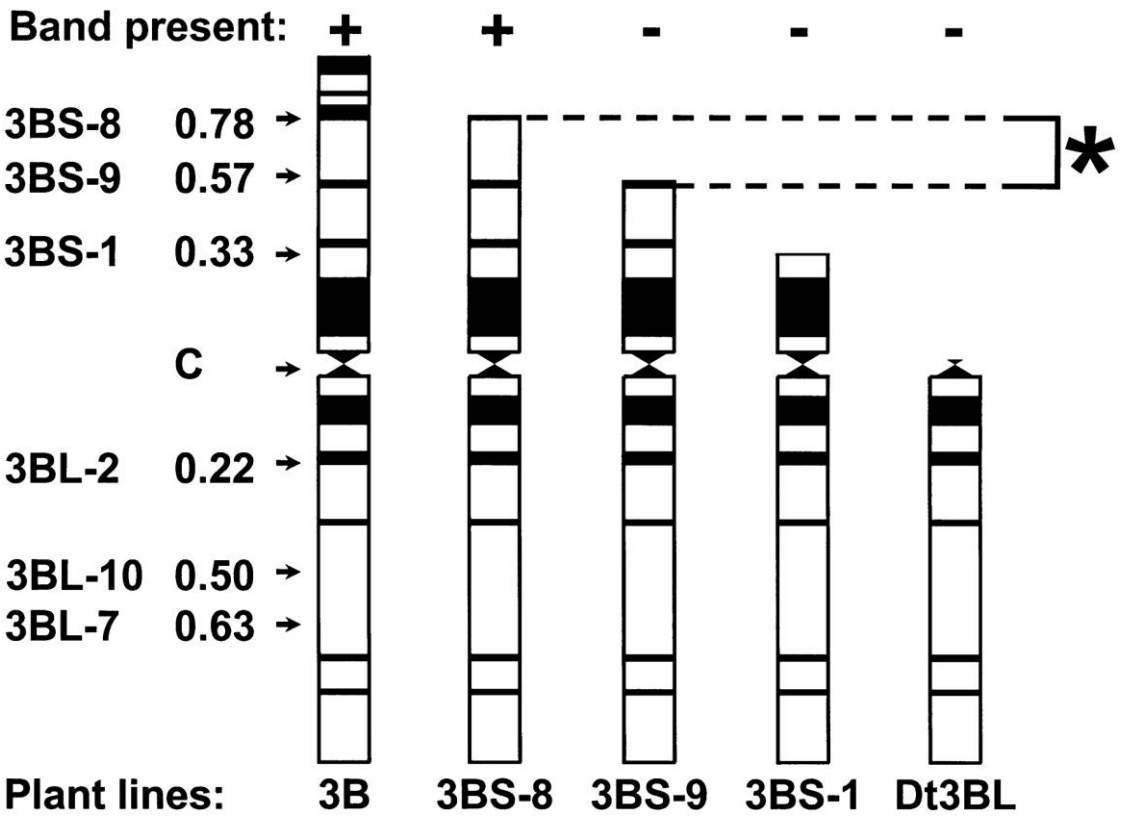


Figure 20 : Exemple de cartographie d'un marqueur dans les bins de délétion. Les cinq figures représentent cinq configurations du chromosome 3B. Les noms des cassures et leur position sur leurs bras de chromosome respectifs sont indiqués sur la gauche et le lettre C désigne le centromère. Le chromosome 3B entier se trouve sur la gauche. Les trois suivants sont des lignées de délétion dont le nom correspond à l'endroit de la délétion. Le dernier (Dt3BL) est la lignée ditélosomique 3BL qui contient une paire de chromosome 3B dont les bras courts sont absents. La présence ou l'absence d'un marqueur dans ces lignées est indiquée par + ou -. La présence du marqueur dans les deux premières lignées mais pas dans les trois dernières indique qu'il est localisé dans le bin 3BS9-0.57-0.78 indiqué par un astérisque. (d'après Lazo et al., 2004)

Finalement, les estimations concernant le contenu de l'espace génique chez le blé sont très variables selon les études. Les sous-génomes du blé contiendraient entre 25000 et 98000 gènes chacun soit une variation du simple au quadruple. Ainsi, la question du nombre de gènes dans le génome du blé fait toujours débat à l'heure actuelle mais à l'avenir, le séquençage de chromosomes individuels devrait aider à affiner son estimation (Choulet et al., 2010).

3.2.2. L'organisation de l'espace génique

Du fait de l'absence de séquence complète du génome du blé, l'organisation de l'espace génique a, jusqu'à présent, dû être estimée à l'aide des ressources génomiques disponibles et des résultats controversés ont été obtenus.

3.2.2.1. La cartographie de marqueurs issus d'ADNc et de clones *PstI*

Malgré la complexité de son génome, des ressources originales ont pu être développées chez le blé. En effet, du fait de sa polyploïdie récente, les pertes de fragments chromosomiques, de bras de chromosome ou même de chromosomes entiers sont tolérées et compensées par les chromosomes homéologues (Qi et al., 2002). Dès les années 50, des lignées aneuploïdes, possédant des nombres anormaux de chromosomes, ont été développées (Sears, 1954; Sears and Sears, 1978). Des lignées nulli-tétrasomiques, pour lesquelles une paire de chromosomes est manquante et est remplacée par une paire de chromosomes homéologues, ont été développées. Plus tard, Endo (1988) a utilisé un chromosome gamétocide d'*Aegilops cylindrica* pour générer des cassures chromosomiques. Par cette technique, des lignées dites de délétion, pour lesquels des fragments jusqu'à des bras entiers de chromosomes sont absents, ont été développées pour chaque chromosome du blé. Les fragments manquants dans ces lignées de délétion ont été caractérisées cytogénétiquement par Endo et Gill (1996). Désormais, ces lignées permettent de cartographier des marqueurs ou même des gènes dans ces fragments de chromosomes, appelés bins de délétion, comme démontré notamment par Lazo et al. (2004) (Figure 20).

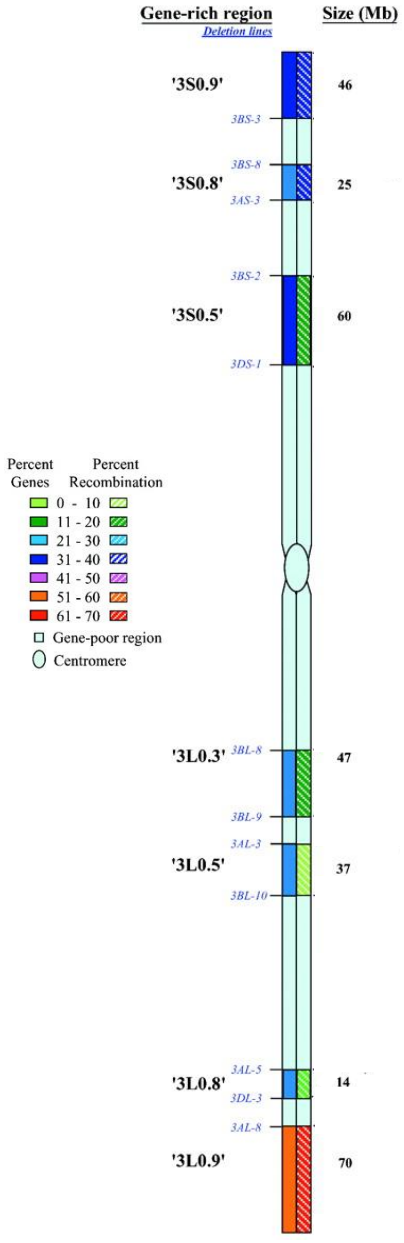


Figure 21 : La distribution des gènes sur les chromosomes du groupe 3. Les noms des régions riches en gènes sont indiqués en noir à gauche du chromosome consensus du groupe 3. Les bins de délétion aux extrémités des régions riches en gènes sont indiqués en bleu à gauche du chromosome. La taille des régions riches en gènes en Mb est donnée à droite du chromosome. Les pourcentages de gènes cartographiés et le taux de recombinaison dans les régions riches en gènes sont représentés en couleur selon la légende (d'après Erayman et al., 2004)

Les premières études concernant l'organisation de l'espace génique ont été réalisées par cartographie de marqueurs dans les bins de délétion par Southern blot (Gill et al., 1996a; Gill et al., 1996b; Sandhu et al., 2001; Sandhu and Gill, 2002; Erayman et al., 2004). Les sondes utilisées pour représenter les gènes étaient essentiellement issues d'ADNc ou de clones génomiques issus de digestion à l'enzyme *PstI*. L'enzyme *PstI* étant sensible à la méthylation, les fragments génomiques générés suite à une digestion utilisant cette enzyme devaient représenter la fraction hypométhylée du génome soit la partie exprimée.

Grâce à cette stratégie et à la cartographie de 82 marqueurs, Gill et al. (1996a) ont constaté que plus de 60% des marqueurs du bras long du groupe de chromosomes 5 étaient présents dans trois clusters représentant moins de 18% de la taille physique du bras. De plus, les régions les plus proximales encadrant les centromères et représentant 20% des chromosomes ne contenaient aucun marqueur. Gill et al. (1996b) ont ensuite cartographié 50 marqueurs dans les bins de délétion du groupe de chromosomes 1 et ont identifié que 86% des marqueurs étaient localisés dans cinq clusters représentant moins de 10% des chromosomes. Sandhu et al. (2001) se sont concentrés sur le bras court du groupe de chromosomes 1 et ont trouvé que 78% des marqueurs étaient cartographiés dans deux régions représentant 14% de la taille du bras. Ainsi, Sandhu et Gill (2002) concluent de ces analyses que chaque chromosome de blé compte entre six et huit régions riches en gènes qui couvriraient environ 10% du chromosome entier. Ils suggèrent également que ces régions riches en gènes sont elles-mêmes constituées de sous-régions riches en gènes et pauvres en gènes représentant 10% et 90% de la région respectivement.

Enfin Erayman et al. (2004) a mené le même type d'expérience mais sur l'ensemble des chromosomes du blé. Ils ont utilisé 942 marqueurs dont 94% ont été cartographiés dans 48 régions riches en gènes représentant au total 29% de la taille du génome (Figure 21). Ces régions seraient également le siège de 95% de la recombinaison. De plus, 80% des marqueurs ont été positionnés dans les moitiés distales des chromosomes et 58% dans les régions les plus distales représentant 20% des chromosomes. En conclusion, d'après ces études, les gènes du blé seraient localisés dans des régions de petite taille essentiellement situées dans les parties distales des chromosomes. En dehors de ces régions représentant un tiers du génome, les gènes seraient quasiment absents et les rétrotransposons formeraient de larges « déserts » (Sandhu and Gill, 2002).

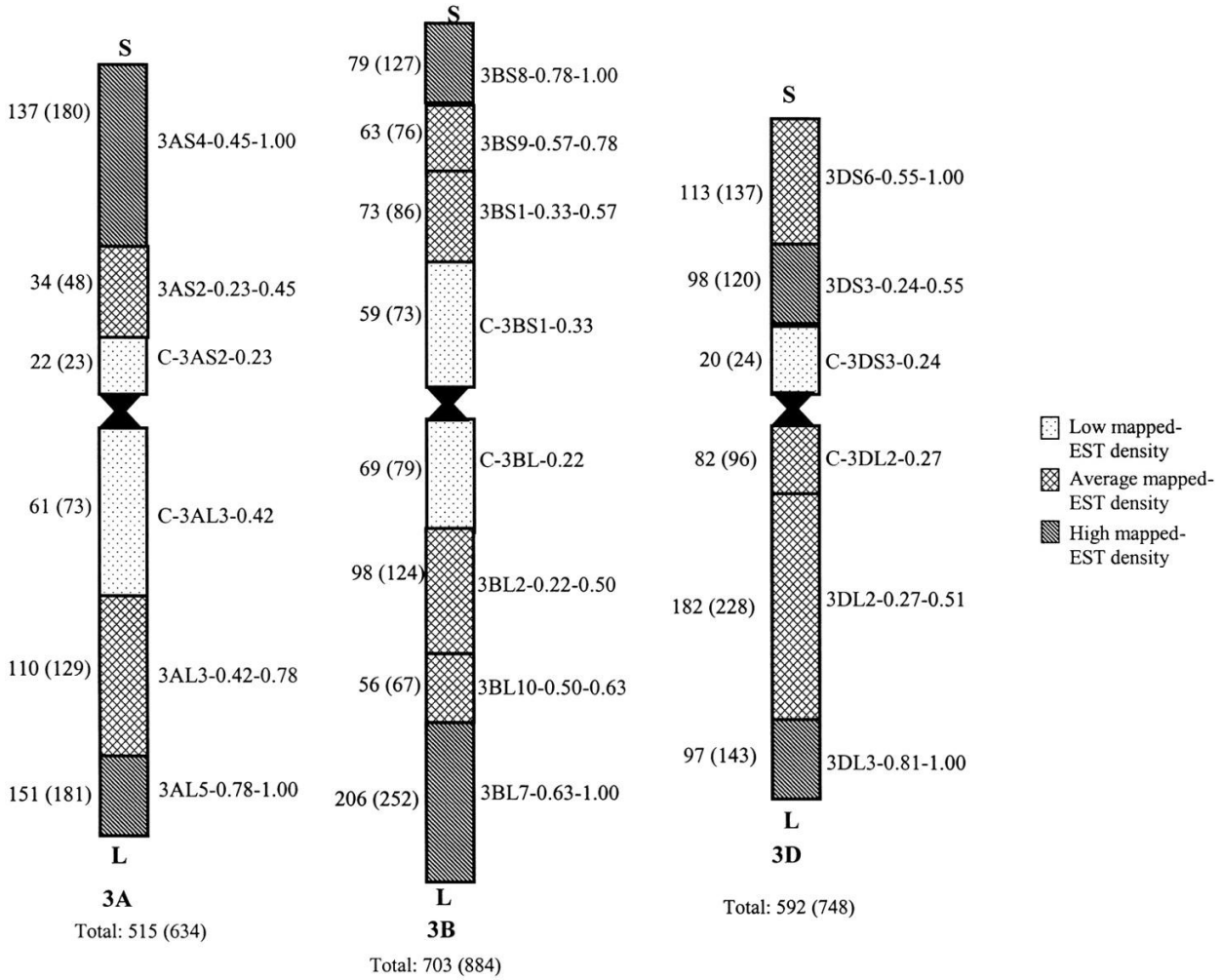


Figure 22 : Représentation graphique des EST cartographiées et des densités d'EST par bin de délétion pour chacun des chromosomes du groupe 3. Les nombres sans parenthèses correspondent au nombre d'EST cartographiées dans les bins. Les nombres entre parenthèses correspondent au nombre de fragments de restriction (locus) cartographiés dans les bins. Les noms des bins sont indiqués à gauche des chromosomes. (d'après Munkvold et al., 2004)

3.2.2.2. La cartographie d'EST

D'autres études ont utilisé la cartographie d'EST dans les bins de délétion par Southern blot pour étudier l'organisation de l'espace génique chez le blé. Grâce à la cartographie de 730 EST le long des chromosomes du blé, Akhunov et al. (2003) ont observé une légère corrélation entre la densité de gènes et la distance au centromère (r = 0,22) mais aussi entre la densité de gènes et le taux de recombinaison (r = 0,22). Ainsi, un gradient positif de la densité de gènes était également observé mais bien moins important que celui suggéré par les études précédemment citées. Aussi, même les parties proximales portaient des EST. Akhunov et al. (2003) expliquent cette différence par le type de marqueurs utilisés qui serait biaisé en faveur des régions télomériques. De plus, ces auteurs ont identifié un gradient positif de densité de gènes dupliqués orienté des centromères vers les télomères.

Par la suite, plus de 7000 EST ont été cartographiées le long des chromosomes du blé (Qi et al., 2004). En général, la densité de gènes le long de chaque chromosome augmentait avec la distance au centromère (r = 0,57). De plus, la plupart des régions pauvres en gènes se situaient dans les bins centromériques. Cependant ces régions n'étaient pas complètement dépourvues d'EST. Des analyses indépendantes des sept groupes de chromosomes de blé ont été réalisées (Conley et al., 2004; Hossain et al., 2004; Linkiewicz et al., 2004; Miftahudin et al., 2004; Munkvold et al., 2004; Peng et al., 2004; Randhawa et al., 2004). Par exemple, 996 EST ont été cartographiés sur le groupe de chromosomes 3 du blé (Munkvold et al., 2004). Pour les chromosomes 3A, 3B et 3D, les bins centromériques étaient les moins denses en EST alors que généralement, les bins télomériques étaient les plus denses (Figure 22). Les parties intermédiaires des bras de chromosomes avaient en général une densité de gènes moyenne. Le bras court du chromosome 3D représente une exception puisque la partie intermédiaire du bras était plus dense en gènes que la partie télomérique. Malgré cette exception, le groupe de chromosomes 3 suivait la tendance générale décrite par Qi et al. (2004) sur l'ensemble des chromosomes. Les bins centromériques des chromosomes du groupe 3 n'étaient pas dépourvus de gènes comme suggéré dans les études précédemment citées mais comptaient entre 16 et 18% des EST du chromosome. Ainsi, grâce à la cartographie d'EST dans les bins de délétion, l'organisation de l'espace génique a été affinée. La densité de gènes serait donc beaucoup plus homogène le long des chromosomes que ce qui avait pu être suggéré au préalable. Elle suivrait un gradient positif des centromères vers les télomères des chromosomes et entre 15 et 20% des gènes seraient localisés au niveau des centromères.

3.2.2.3. Le séquençage et l'annotation de BAC

De plus, l'obtention de séquences génomiques et leur annotation a permis d'apprécier l'organisation de l'espace génique du blé à une échelle beaucoup plus fine. Dès 1999, Feuillet et Keller (1999) ont analysé l'organisation de l'espace génique du blé à partir d'une séquence de 13,9 kb qui contenait un îlot de trois gènes de résistance. La densité de cette région était de un gène par 4,6 kb. De plus, la région orthologue chez l'orge a également été séquencée et portait cinq gènes dans 23 kb soit une densité identique à celle calculée sur la séquence de blé. Ainsi, cette étude a mis en évidence que des régions ayant des densités en gènes similaires à celle observée dans le génome d'*Arabidopsis* existent dans les grands génomes du blé et de l'orge. Par la suite, Brooks et al. (2002) ont séquencé un BAC de blé entier de 106 kb porteur d'un gène de résistance. Ce BAC portait 12 gènes au total et avait une densité de un gène tous les 8,9 kb. Cependant cette densité ne reflète pas la distribution hétérogène des gènes le long de cette séquence. En effet, un îlot de sept gènes a été identifié dans 46 kb soit une densité de un gène par 6,6 kb puis les 30 kb suivants correspondaient à des ET et enfin cinq gènes ont été annotés en 19 kb soit une densité de un gène par 3,8 kb. Ainsi, une séquence plus longue a permis de mettre en évidence une alternance entre des îlots de gènes dont la densité est similaire à celle d'*Arabidopsis* et des régions composées exclusivement d'ET. Une telle organisation a été confirmée par Salina et al. (2009) qui ont séquencé un BAC du génome B du blé porteur d'une séquence subtélomérique d'*Aegilops*. Ce BAC de 120 kb portait cinq gènes soit une densité de un gène par 24 kb mais ces derniers étaient localisés sur 33 kb et formaient un îlot d'une densité de un gène par 6,6 kb. Enfin, le séquençage d'un BAC de *Triticum monococcum* de 101 kb porteur du locus *Ha* a été réalisé par Chantret et al. (2004). Comme sept gènes ont été annotés, la densité de ce BAC était de un gène tous les 14 kb. Cependant, un îlot de cinq gènes a été annoté en 27 kb soit une densité de un gène par 5,4 kb puis un gène a été trouvé isolé au milieu de 69 kb d'ET et le dernier gène a été positionné en fin de séquence. Cette dernière étude montre ainsi que tous les gènes ne sont pas nécessairement organisés en îlots mais qu'ils peuvent également être isolés au milieu des ET.

Cependant ces quatre études ont été réalisées sur des BAC sélectionnés au niveau des télomères ou pour les gènes de résistance ou d'intérêt agronomique dont ils étaient porteurs. Cette sélection ayant pu entraîner un biais en faveur des régions très denses (Devos et al., 2005; Breen et al., 2010), des analyses de BAC choisis aléatoirement ont été réalisée. Devos et al. (2005) ont séquencés quatre BAC dont trois portaient un gène et le dernier portait deux

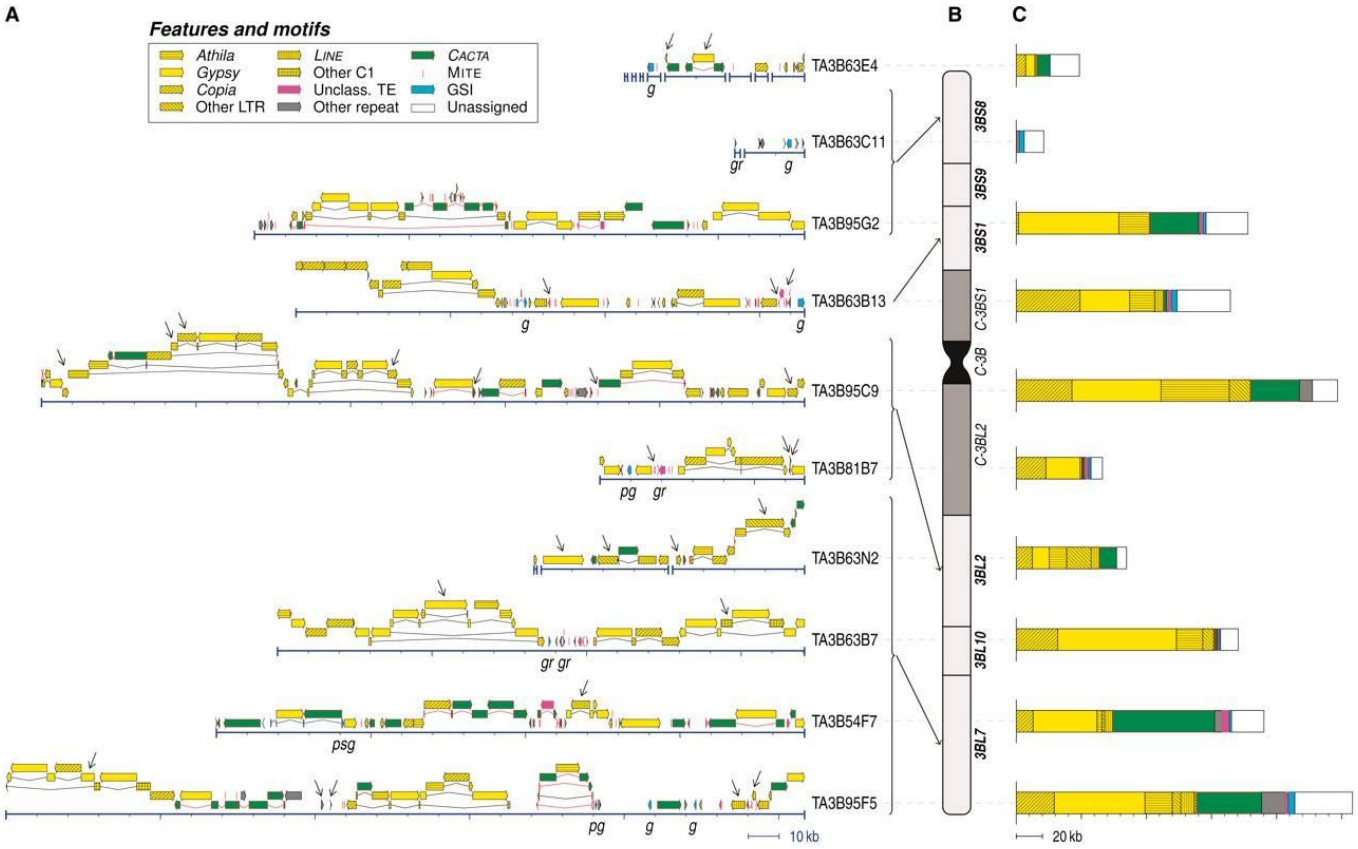


Figure 23 : Annotation détaillée, position dans les bins et composition des dix BAC séquencés du chromosome 3B du blé. (A) Les annotations détaillées des dix BAC séquencés. Les principaux ET, les autres séquences répétées et les gènes (GSI) sont représentés comme indiqués par la légende « Features and motifs ». g, gènes ; pg, gènes putatifs ; gr, reliques de gènes ; psg, pseudogènes. Pour les imbrications d'ET, les ET les plus récemment insérés sont représentés au-dessus des plus anciens représentés sous forme découpée. (B) Position de neuf BAC dans les bins de délétion. (C) Proportions des principaux constituants des BAC. (d'après Charles et al., 2008)

gènes. La densité globale était de un gène par 83 kb. Pourtant trois BAC ont été positionnés dans des régions riches en gènes alors que leur densité n'était que de un gène tous les 75 kb. Cette étude permet de constater que les régions denses en gènes du génome du blé ne sont pas forcément aussi denses que ce qui avait été suggéré par les analyses de BAC ciblés. Cependant une densité de un gène tous les 75 kb reste deux fois plus importante que la densité moyenne attendue de un gène par 113 kb correspondant aux 150000 gènes répartis aléatoirement le long des 17 Gb de séquence du génome du blé. Une autre étude menée sur dix BAC choisis dans la banque BAC du chromosome 3B du blé confirme cette observation (Charles et al., 2008). Treize gènes ont été annotés sur ces dix BAC couvrant 1,43 Mb au total, soit une densité de un gène tous les 110 kb (Figure 23). Pourtant trois BAC ne portaient pas de gènes et n'étaient pas localisés dans les régions centromériques du chromosome. Ainsi, le séquençage de BAC sélectionnés au hasard suggère une alternance entre les îlots de gènes, les gènes isolés et les océans d'ET mais avec des îlots de gènes généralement moins denses que ce qui avait été constaté avec des BAC ciblés.

Enfin, très récemment, ces analyses ont été complétées par l'acquisition de séquences issues de BAC chevauchants formant des séquences contigües plus longues. Breen et al. (2010) ont séquencé et analysé un contig de 784 kb composés de 15 BAC chevauchants au sein duquel un îlot de trois gènes a été annoté. La densité en gènes de ce contig est de un gène par 260 kb alors que les trois gènes sont localisés dans une portion de séquence de 14 kb soit une densité locale de un gène par 4,6 kb. Le reste du contig est exclusivement constitué de rétrotransposons à LTR insérés les uns dans les autres. Pourtant ce contig est positionné dans le bin de délétion le plus télomérique du bras long du chromosome 3B supposé dense en gènes. Les auteurs proposent donc que ce contig serait localisé au niveau d'une des deux petites régions hétérochromatiques identifiées dans ce bin de délétion. De plus, Choulet et al. (2010) ont analysé les séquences de 13 contigs de BAC issus de différentes régions du chromosome 3B soit un total de 18,2 Mb. Comme 175 gènes ont été annotés, la densité globale était de un gène par 104 kb mais de fortes disparités ont été observées en fonction des contigs. En effet, les contigs localisés en position distale étaient deux fois plus denses en gènes (un gène par 86 kb) que les contigs en position proximale (un gène par 184 kb), suggérant un gradient de la densité de gènes le long du chromosome 3B. Malgré cette différence, comme 73% des BAC portaient des gènes, Choulet et al. (2010) estiment que les régions de plusieurs mégabases sans gènes doivent être rares. A une échelle plus fine, des

disparités dans la répartition des gènes ont également été constatées. En effet, les distances intergéniques observées étaient très variables avec une moyenne de 96 ± 128 kb. Elles sont essentiellement constituées d'ET insérés les uns dans les autres. Cependant 50% des distances intergéniques comptaient moins de 43 kb. En utilisant cette valeur comme seuil pour définir les îlots de gènes, Choulet et al. (2010) ont identifié que 75% des gènes étaient organisés en îlots de moins de quatre gènes.

Finalement, les analyses de séquences ont permis d'affiner notre compréhension de l'organisation hétérogène de l'espace génique à l'échelle des gènes. Elles suggèrent une distribution des gènes en îlots plus ou moins denses séparés par des océans d'ET eux-mêmes interrompu occasionnellement par des gènes isolés. Les analyses récentes de Choulet et al. (2010) confirment également le gradient positif de gènes des centromères vers les télomères identifiés par cartographie d'EST. En outre, à partir de données non publiées, Devos (2010) relierait les observations réalisées au niveau de la séquence avec le gradient de gènes le long des chromosomes de blé. En effet, les distances entre îlots diminueraient des centromères vers les télomères alors que le nombre de gènes par îlot augmenterait.

En conclusion, l'organisation de l'espace génique est une question qui préoccupe la communauté scientifique du blé depuis presque 15 ans. Les résultats initiaux proposant une répartition des gènes exclusivement dans les parties télomériques extrêmes ont été nuancés au gré des améliorations technologiques. Ainsi, la synthèse des résultats des études les plus récentes suggère que les gènes seraient répartis de façon hétérogène le long des chromosomes de blé. Leur densité serait plus importante au niveau des télomères qu'au niveau des centromères. A une échelle plus fine, une alternance entre des îlots de gènes, des océans d'ET et des gènes isolés serait observable.

PRESENTATION DU PROJET DE LA THESE

Mon travail de thèse avait donc pour objectif d'approfondir notre connaissance de l'espace génique du blé en tentant notamment d'étudier l'impact de la structure du génome (polyploïde, de grande taille et riche en séquences répétées) sur l'organisation, la régulation et la fonction des gènes sur le chromosome 3B du blé hexaploïde.

Grâce au développement de nouvelles ressources génomiques originales au sein de mon équipe d'accueil « Structure, fonction et évolution des génomes de blé » dirigée par Catherine Feuillet à l'INRA de Clermont-Ferrand, j'ai pu développer une stratégie innovante à différentes échelles pour analyser et comprendre l'organisation de l'espace génique chez le blé tendre.

Les outils de génomique ont été développés grâce à une approche originale initiée par l'équipe visant à réduire la complexité de l'analyse du génome du blé en étudiant les chromosomes individualisés. Cette approche chromosome-spécifique a été rendue possible par la mise au point d'une technologie permettant d'isoler des chromosomes ou des bras de chromosomes par cytométrie de flux (Safar et al., 2004). Le chromosome 3B du blé, dont la taille (1 Gb) est significativement supérieure à celle des autres chromosomes, a été le premier isolé par cette technique et a fait l'objet d'un projet pilote mené par l'équipe. Tout d'abord, la pureté ainsi que la quantité de chromosomes 3B isolés par cette technique ont permis l'élaboration d'une banque BAC spécifique de ce chromosome. Celle-ci contient environ 68000 clones BAC d'une taille moyenne de 103 kb et représente une couverture équivalente à 6,2 x du chromosome 3B (Safar et al., 2004). De plus, l'élaboration de la banque BAC spécifique du chromosome 3B a permis à l'équipe de construire la première carte physique d'un chromosome de blé (Paux et al., 2008). Elle compte 1036 contigs couvrant 82% du chromosome 3B. Les deux tiers de ces contigs soit 611 Mb sont ancrés sur la carte génétique du chromosome 3B à l'aide de 1443 marqueurs. De plus, 556 Mb ont été cartographiés dans les bins de délétion, grâce aux 16 lignées de délétion disponibles pour le chromosome 3B.

Grâce à ses ressources uniques développées chez le blé, j'ai pu mettre en œuvre une stratégie originale basée sur des outils d'analyse du transcriptome pour étudier l'organisation de l'espace génique du blé à l'échelle du chromosome 3B entier avec une résolution à l'échelle du BAC. Cette approche m'a permis d'appréhender l'organisation de l'espace génique de la façon la plus exhaustive jamais réalisée à l'heure actuelle sur un chromosome de blé et à une résolution plus fine que celle du bin de délétion précédemment utilisée. La variabilité de la

densité de gènes le long du chromosome 3B a été plus particulièrement examinée pour vérifier l'hypothèse du gradient positif du centromère vers les télomères. De plus, la variabilité de la répartition des gènes à l'échelle du BAC a été analysée afin d'évaluer la proportion de gènes organisés en îlots. Cette étude approfondie de l'organisation de l'espace génique le long d'un chromosome de blé s'est également inscrite dans le choix d'une stratégie de séquençage des chromosomes de blé.

Après l'analyse de l'organisation de l'espace génique le long du chromosome, j'ai mené une analyse fonctionnelle des gènes portés par le chromosome afin d'identifier des régions chromosomiques ayant des particularités du point de vue de l'expression, de la régulation et de la fonction des gènes. De plus, à l'échelle du BAC, les caractéristiques fonctionnelles ont été analysées afin d'étudier leur impact sur les répartitions de gènes observées. Enfin, l'organisation de l'espace génique a également été analysée d'un point de vue évolutif à l'échelle du chromosome et du BAC pour tenter de comprendre comment cette structuration a été orchestrée et mise en œuvre au cours de l'évolution du génome du blé tendre.

RESULTATS DE LA THESE

Dans un premier temps, mon travail de thèse a consisté à appréhender l'organisation de l'espacé génique à l'échelle du chromosome 3B entier de la façon la plus exhaustive possible compte tenu des ressources génomiques disponibles. Pour cela, une approche de transcriptomique a été développée pour cartographier la fraction transcrite du génome sur les BAC formant le « Minimal Tiling Path » (MTP) ou chemin de chevauchement minimal couvrant 82% du chromosome 3B. Ce travail a été mené en parallèle avec un projet de séquençage et d'annotation de séquences de contigs de plusieurs mégabases issues du chromosome 3B mené par Frédéric Choulet. Bien que la résolution de mon étude soit très éloignée de l'analyse de ces contigs, mes résultats obtenus à l'échelle du chromosome 3B entier apportent des éléments de confirmation aux tendances observées à l'échelle de séquences ne couvrant que 2% du chromosome et permettent d'extrapoler certaines de ces observations à l'ensemble du chromosome. Les résultats que j'ai obtenus ont été associés à ceux de Frédéric Choulet et publiés dans la revue Plant Cell sous le titre « Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces » (Choulet et al., 2010).

ARTICLE N°1

L’espace génique couvre l’ensemble du chromosome 3B sans aménager de grandes régions totalement dépourvues de gènes

This article is a *Plant Cell* Advance Online Publication. The date of its first appearance online is the official date of publication. The article has been edited and the authors have corrected proofs, but minor changes could be made before the final version is published. Posting this version online reduces the time to publication by several weeks.

Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces [W]

Frédéric Choulet,[a] Thomas Wicker,[b] Camille Rustenholz,[a] Etienne Paux,[a] Jérome Salse,[a] Philippe Leroy,[a] Stéphane Schlub,[c] Marie-Christine Le Paslier,[c] Ghislaine Magdelenat,[d] Catherine Gonthier,[d] Arnaud Couloux,[d] Hikmet Budak,[e] James Breen,[f] Michael Pumphrey,[g] Sixin Liu,[h] Xiuying Kong,[i] Jizeng Jia,[i] Marta Gut,[j] Dominique Brunel,[c] James A. Anderson,[h] Bikram S. Gill,[g] Rudi Appels,[f] Beat Keller,[b] and Catherine Feuillet[a,1]

[a] Institut National de la Recherche Agronomique, Université Blaise Pascal, Unité Mixte de Recherche 1095 Genetics Diversity and Ecophysiology of Cereals, F-63100 Clermont-Ferrand, France
[b] Institute of Plant Biology, University Zurich, 8008 Zurich, Switzerland
[c] Institut National de la Recherche Agronomique, Unité de Recherche1279 Etude du Polymorphisme des Génomes Végétaux, Commissariat à l'Energie Atomique-Institut de Génomique-Centre National de Génotypage, F-91057 Evry, France
[d] Génoscope, Institut de Génomique, Commissariat à l'Energie Atomique, F-91057 Evry, France
[e] Sabanci University, Engineering and Natural Sciences, Biological Science and Bioengineering Program, 34956 Tuzla-Istanbul, Turkey
[f] Centre for Comparative Genomics, Murdoch University, Western Australia, WA 6150, Australia
[g] Department of Plant Pathology, Wheat Genetic and Genomic Resources Center, Kansas State University, Manhattan, Kansas 66506-5502
[h] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, Minnesota 55108
[i] National Key Facility for Crop Gene Resources and Genetic Improvement, Key Laboratory of Crop Germplasm Resources and Utilization, Ministry of Agriculture Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R. China
[j] Department of Technology Development, Commissariat à l'Energie Atomique-Institut de Génomique-Centre National de Génotypage, F-91057 Evry, France

To improve our understanding of the organization and evolution of the wheat (*Triticum aestivum*) genome, we sequenced and annotated 13-Mb contigs (18.2 Mb) originating from different regions of its largest chromosome, 3B (1 Gb), and produced a 2x chromosome survey by shotgun Illumina/Solexa sequencing. All regions carried genes irrespective of their chromosomal location. However, gene distribution was not random, with 75% of them clustered into small islands containing three genes on average. A twofold increase of gene density was observed toward the telomeres likely due to high tandem and interchromosomal duplication events. A total of 3222 transposable elements were identified, including 800 new families. Most of them are complete but showed a highly nested structure spread over distances as large as 200 kb. A succession of amplification waves involving different transposable element families led to contrasted sequence compositions between the proximal and distal regions. Finally, with an estimate of 50,000 genes per diploid genome, our data suggest that wheat may have a higher gene number than other cereals. Indeed, comparisons with rice (*Oryza sativa*) and *Brachypodium* revealed that a high number of additional noncollinear genes are interspersed within a highly conserved ancestral grass gene backbone, supporting the idea of an accelerated evolution in the *Triticeae* lineages.

INTRODUCTION

Bread wheat (*Triticum aestivum*) is allohexaploid with three homoeologous genomes (2n=6x=AABBDD) and has one of the largest higher plant genomes (17 Gb, 40 times larger than the rice (*Oryza sativa*) genome [International Rice Genome Sequencing Project, 2005]). Because of its size and high repetitive sequence content (~80%; Smith and Flavell, 1975), sequencing the wheat genome has been perceived as too challenging; consequently, its organization and composition remain largely unknown. Most of the wheat genomic sequences available to date have been obtained during map-based cloning projects or comparative studies at disease resistance, storage protein, grain hardness, domestication, or vernalization loci (for review, see Feuillet and Salse, 2009; Krattinger et al., 2009). In all cases, the sequences corresponded to single BAC clones or small BAC contigs of 129 kb on average, and only five contigs larger than 300 kb are available to date in the databanks (http://srs.ebi.ac.uk), the largest of which has a maximum contiguous size of 450 kb

[1] Address correspondence to catherine.feuillet@clermont.inra.fr.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Frédéric Choulet (frederic.choulet@clermont.inra.fr).
[W] Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.110.074187

(Ogihara et al., 2005). Analyses of the 3.8 Mb representing all wheat genomic sequences available in the public databases in 2005 showed an average gene density of 1 gene per 24 kb with only ~55% of transposable elements, thereby indicating a clear bias toward gene-rich regions in these samples (Sabot et al., 2005). Early studies based on EST mapping in cytogenetic bins suggested that the gene distribution is highly heterogeneous in wheat with >90% of the gene space clustered in 29% of the genome, mainly in the telomeric part of the chromosomes (Erayman et al., 2004; Qi et al., 2004). By contrast, sampling of the wheat genome performed through the sequencing of randomly chosen BAC clones or large samples of BAC end sequences (BES) provided evidence for a more homogeneous gene distribution in the genome. In a preliminary study of four BAC clones from a whole bread wheat genome (cv Chinese Spring) BAC library, Devos et al. (2005) found an average gene density of 1/75 kb, while Charles et al. (2008) reported a gene density of 1/100 kb homogeneously distributed after sequencing 10 BACs (1.43 Mb) randomly chosen from the 3B chromosome-specific BAC library. Finally, end sequencing of 10,000 BAC clones distributed along the physical map of chromosome 3B led to an estimate of 1 gene/165 kb (Paux et al., 2006). Recently, draft sequences of 217 additional BAC clones from the Chinese Spring BAC library have been deposited in the databanks (AC200765-851, AC207901-60, AC216550-85, AC232247-62, AC238983-88, and DQ767609-30) by Bennetzen et al., and annotation to refine these estimates is underway (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0501814).

Further insights into the composition of the wheat genome have been provided by other random sequencing efforts. Analysis of 3 Mb of plasmid end sequences produced from *Aegilops tauschii* (the D genome progenitor of hexaploid bread wheat) indicated that 92% of repeated sequences (Li et al., 2004) contained 68% of transposable elements (TEs), whereas the annotation of 11 Mb of BESs from the 3B chromosome of hexaploid wheat (Paux et al., 2006) revealed a repeat content of 86% as well as an estimated gene number of 6000 for chromosome 3B (i.e., 36,000 genes per diploid genome). This contrasted dramatically with the predictions of Rabinowicz et al. (2005) who suggested as much as 98,000 genes per subgenome in bread wheat based on the analysis of 1597 plasmid ends from a methyl-filtration library. Similar variations in gene number were reported for maize (*Zea mays*) until its genome was fully sequenced (Schnable et al., 2009). Early genetic analyses suggested that maize genes are clustered primarily in high-coding density islands representing 10 to 20% of the genome (Carels et al., 1995). Sequencing of large sets of random BAC clones led to the revision of the gene island concept and suggested that the 42,000 to 59,000 estimated genes are largely spread in 78% of the genome (Haberer et al., 2005). Finally, the first improved maize genome sequence draft provided an estimate of 32,000 non-TE-related genes (Schnable et al., 2009) that are found with an increased density in the subtelomeric regions (Soderlund et al., 2009), a pattern that was also observed in sorghum (*Sorghum bicolor*; Paterson et al., 2009).

During its evolution, the hexaploid wheat genome has been shaped by different evolutionary forces. At the global level, its basic structure results from large rearrangements that range from an ancestral grass whole-genome duplication, to chromosome fusions and chromosome number reduction (Luo et al., 2009; Salse et al., 2009), and more recently to two polyploidization events that brought together the A, B, and D genomes into a single nucleus (McFadden and Sears, 1946). At the sequence level, its organization and composition can be depicted as two main compartments with different evolutionary dynamics and relative importance: a small conservative part that is subjected to selection pressure and mostly corresponds to the gene space and a much larger and more variable component that is under more dynamic evolution and comprises the TE space as well as duplicated genes and gene fragments. Evidence for some of the mechanisms underlying the dynamics of these two compartments (e.g., TE insertions, illegitimate and unequal recombination, and interchromosomal and tandem duplications) have been provided by comparative analyses of homeologous loci (for review, see Feuillet and Salse 2009) as well as by gene family studies (Akhunov et al., 2007), but, to date, little is known about the extent and relative impact of these mechanisms on the organization and distribution of the gene and TE spaces along the wheat chromosomes.

The current lack of knowledge about the complexity and organization of the hexaploid wheat genome sequence hampers the delineation of the most cost-effective and informative sequencing strategy even with the advent of the Next Generation Sequencing technologies (Metzker, 2009). To reduce the complexity of the analyses, the International Wheat Genome Sequencing Consortium (www.wheatgenome.org) embarked a few years ago on a chromosome-based approach (Dolezel et al., 2009). Recently, we provided a proof of concept for this approach with the construction of the physical map of the 1 Gb wheat chromosome 3B (Paux et al., 2008). In this study, we gained additional knowledge about the organization and composition of the wheat genome by the complete sequencing and annotation of 13 Mb-sized (0.3 to 3.1 Mb) BAC contigs selected from different regions of the chromosome and by whole chromosome shotgun sequencing. Our data show that (1) genes are present along the whole chromosome and are clustered mainly into numerous, very small islands separated by large blocks of repetitive elements, and (2) genome expansion occurred homogeneously along the chromosome through specific TE bursts. In addition, they reveal an accelerated evolution through tandem or interchromosomal gene duplications in the telomeric regions that led to an increase in the gene number in wheat compared with related grasses without disruption of the ancestral gene backbone. These gene rearrangements combined with the differential insertion or removal of specific TE families resulted in a contrasted sequence composition that is now observed between the proximal and distal regions of the wheat chromosomes.

RESULTS

Mb-Sized Contig and Whole Chromosome 3B Shotgun Sequencing

Thirteen contigs representing 152 BAC clones were selected for sequencing out of the 1036 established for the physical map of chromosome 3B (Paux et al., 2008). Twelve contigs originated

from seven different deletion bins on the short (four contigs) and long (eight contigs) arms of the chromosome, while one contig was from the centromere (Table 1). Five of the long-arm contigs are part of the large (200 Mb) telomeric bin 3BL7-0.63-1.00 and were selected for their putative contrasted gene content estimated after screening of the 3B BAC library with a set of 399 ESTs previously assigned to this bin (see Methods). On the short arm, two contigs (*ctg0954* and *ctg0011*) were selected from a region of 12 centimorgans located between the markers *gwm389* and *gwm493* that carries a high density of disease resistance genes (Paux et al., 2008). The 152 BAC clones were sequenced by Sanger or 454 Roche GSFLX technologies and completely assembled into single scaffolds. This resulted in 18.212 Mb, including eight contiguous sequences larger than 1 Mb (up to 3.1 Mb) that represent 2% of the 3B chromosome and 0.3% of the whole B genome (accession numbers FN564426-37 and FN645450). To support sequence annotation and obtain additional whole 3B chromosome sequence data, Illumina/Solexa sequencing was performed on sorted and amplified 3B chromosomes. A total of 54,808,646 short reads of 36 bp were generated, resulting in 1,973,111,256 nucleotides that represent 2X coverage of the chromosome (accession number ERA000182). This was used to establish a Mathematically Defined Repeats (MDR) index by counting occurrence of 17-mers and to evaluate the number of genes carried by chromosome 3B. High-quality sequence annotation was performed with a combination of automated procedures and manual curation based on sequence similarities and MDR patterns (see Methods).

General Features of the Sequence Composition

Annotation of the 18.2 Mb of contig sequences revealed 199 non-TE genic features that were classified into three categories (see Methods for definitions): 148 protein coding genes, 27 pseudogenes, and 24 gene fragments (Table 1; see Supplemental Table 1 online for gene functions). The gene assignments were all supported by at least a full-length cDNA, an EST, and/or a homolog in another genome. Among these assignments, 76% showed a hit with one of the 40,349 *T. aestivum* unigenes (National Center for Biotechnology Information [NCBI] build#55). The coding fraction of the sample represents 242 kb (i.e., 1.5%) of the sequences, while the TE content is 81.4% (graphical view of the annotations in Supplemental Figure 1 online). An average GC content of 46.2% was found for all BACs (SD =1.6%), whereas the 1973 Mb of Solexa reads generated from sorted chromosomes and previous analysis of 11 Mb of BES (Paux et al., 2006) indicated significantly lower values of 42.8 and 44.5%, respectively. In contrast with the constant GC content, the composition in genes and TEs was highly variable between the different BAC sequences. The proportion of TEs ranged from 19 to 100%, while the gene content ranged from 0 to 10 genes per BAC. Coding exons show a high GC content: 59.0% ± 8.5%, which decreases to 54.7 ± 8.1 for pseudogenes and to 51.5 ± 9.5 for gene fragments, indicating that loss of function is followed by a period of relaxed selection for codon usage.

Gene Structures and Intron-Associated TEs

Analysis of the 175 gene and pseudogene models revealed a gene size ranging from 309 bp to 15.8 kb with an average of 3300 ± 2900 bp, an average number of 5.6 ± 5.5 exons per gene (median = 4), and 20% of genes without introns (see Supplemental Figure 2 online). One-third of the genes contained only one or two exons, whereas,18% had 10 or more exons comparable to what has been described for maize (Haberer et al., 2005). The average coding sequence (CDS) size of 1382 ± 852 bp

Table 1. Features of the 13-Mb Contig Sequences from Chromosome 3B

Bin	Contig Name	Contig Size (bp)	No. BACs	No. Genes	No. Pseudogenes	No. Gene Fragments	Gene Density (kb^{-1})	No. Genes per Mb	TE Content (%)	GC Content (%)
3BS8-0.78-0.87	ctg0011	1266078	16	21	5	10	48.7	20.5	49.7	44.8
3BS8-0.78-0.87	ctg0954	3109948	26	41	8	4	63.5	15.8	63.2	45.6
3BS1-0.33-0.55	ctg1030	619476	6	0	0	0	<619.5	0.0	97.8	48.8
C-3BS5-0.07	ctg1035	711534	5	1	1	0	355.8	2.8	89.9	45.6
Centromere	100L17	268551	1	0	0	0	<268.6	0.0	96.1	43.0
3BL2-0.22-0.28	ctg0616	786544	6	6	0	0	131.1	7.6	90.7	46.7
3BL2-0.22-0.28	ctg0382	1610902	12	9	4	0	123.9	8.1	88.7	46.0
3BL1-0.31-0.38	ctg0005	1715514	12	9	1	1	171.6	5.8	92.1	46.1
3BL7-0.63-1.00	ctg0528	1033236	8	4	0	0	258.3	3.9	91.2	46.3
3BL7-0.63-1.00	ctg0464	2543369	21	22	3	7	101.7	9.8	82.4	46.9
3BL7-0.63-1.00	ctg0091	2776447	23	16	4	0	138.8	7.2	89.4	46.4
3BL7-0.63-1.00	ctg0079	1305738	10	13	0	0	100.4	10.0	88.0	46.4
3BL7-0.63-1.00	ctg0661	465250	6	6	1	2	66.5	15.0	74.5	45.5
		18212587a	152[a]	148[a]	27[a]	24[a]	104.1[b]	9.6[b]	81.5[b]	46.2[b]

Contig localization in deletion bins is indicated from the top to the bottom of the chromosome. The contig size, number of BACs per contig, as well as the number gene, pseudogenes, and gene fragments are indicated. Only genes and pseudogenes were taken into account for calculating the gene density and the number of genes per megabase.

[a]Sum of the column values.

[b]Average gene density (in kb^{-1} and gene per Mb), TE, and GC contents (in %) calculated by considering the sizes of the 13 contigs relative to the total contig length (18 Mb).

(N50 = 1260 bp) was close to the average value (1143 bp) obtained from 6137 full-length cDNAs from wheat (Mochida et al., 2009), indicating that our sample provides a good representation for the wheat genes. Intron sizes were highly variable with a median size of 130 bp, similar to that observed in rice (median 138 bp; Yu et al., 2002), and slightly lower than maize introns (median 166 bp; Haberer et al., 2005). In total, 118 TEs (or truncated derivatives) were found in introns for 33% of the genes. Most of them (82) corresponded to miniature inverted-repeat transposable elements (MITEs), and there was a clear correlation between gene islands and MITE islands along the contigs (see Supplemental Figure 3 online), thereby confirming the preferential association of these small elements with genes (Wessler et al., 1995; Sabot et al., 2005). In addition, long interspersed nuclear elements (LINEs) (*Stasy* family) were also preferentially associated with genes similar to what was observed in rice (International Rice Genome Sequencing Project, 2005) and grapevine (*Vitis vinifera*; Jaillon et al., 2007). Very few complete LINEs (10/170) were found in the data set, confirming that retrotransposition of these elements often leads to the insertion of a small partial element by aborted reverse transcription (Jurka, 1997). By contrast, long terminal repeat (LTR) retrotransposons, which are the most represented TEs in the wheat genome, were almost completely excluded from the genes with only eight LTR retrotransposon fragments detected in introns.

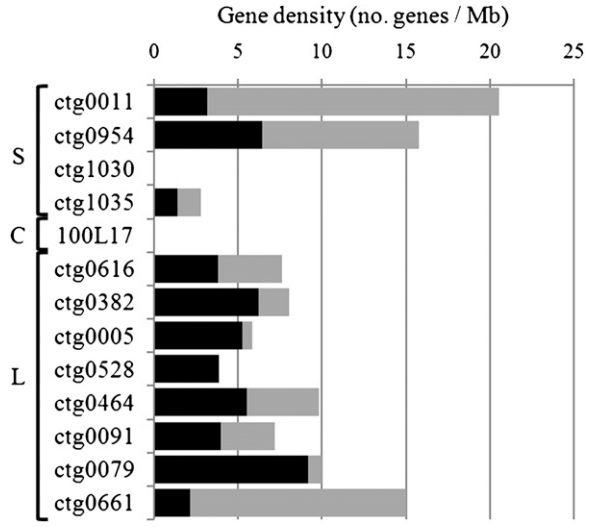


Figure 1. Gene Density and Level of Synteny along the 13 Contigs.

Gene density calculated for the 13 sequenced contigs displayed according to their chromosomal location from the top to the bottom (S, short arm; C, centromeric region; L, long arm). Densities of the syntenic (black) and locus-specific (gray) genes are represented and expressed in number of genes per megabases.

Gene Density and Gene Distribution along Chromosome 3B

The average gene density was 1 gene per 104 kb. Interestingly, comparisons between the distal and proximal contig sequences revealed a twofold increase in gene density toward the telomeres with 1 gene per 184 kb in the proximal regions (*ctg1030*, *1035*, *100L17*, *0616*, *0382*, and *0005*) versus 1 per 86 kb in the more distal regions (*ctg0011*, *0954*, *0528*, *0464*, *0091*, *0079*, and *0661*), thereby suggesting differential gene density distribution along chromosome 3B (Figure 1). At the contig scale, genes were found in 95% of the sample sequences (11/13 contigs), with the exception of two contigs originating from the centromeric region (*100L17*) and the middle of the short arm (*ctg1030*) that both exhibited a TE content of 97% (Table 1). Twenty-eight blocks of TEs larger than 200 kb and representing 8.77 Mb in total were found, the largest of which was 709 kb (in *ctg0091*). Thus, our data indicate that genes are present in the majority (73%) of the BAC clones and that Mb-sized regions without genes are rare. This observation was further supported by hybridization experiments of macroarrays, comprising the 7440 BACs of the 3B physical map minimal tiling path (MTP) (Paux et al., 2008), with 13 different wheat cDNA samples originating from mRNAs extracted at different growth stages and organs. Although it was not possible to determine the exact number of genes present on a BAC with this method, the results showed that 48% (3563) of the MTP BACs carry at least one expressed gene. This ratio is slightly lower than the prediction from the contig sequence annotation, probably since it is based solely on gene expression and is limited by the sensitivity of hybridizations on macroarrays. By combining the hybridization results with the position of each BAC within contigs, we found on average at least one expressed gene every 220 kb. The largest region without detected genes was 840 kb, which is in the range of the 709 kb calculated on the sequenced contigs. In addition, no significant differences in the density of positive BAC clones were observed between the different bins (χ^2 test, P value = 0.39), confirming that genes are distributed across the entire chromosome 3B.

Previous work suggested that genes in wheat are preferentially found in gene islands, but the length and density of such islands had not been defined so far primarily because random single BAC analyses are limited in the ability to estimate average distances between genes. Here, we took advantage of having access to large and contrasted regions to better define the proportion and features of gene islands by calculating first the size distribution of intergenic intervals (Figure 2). One hundred and seventy six intergenic distances (IGDs) were calculated on the basis of the 175 genes and pseudogenes identified within the 13 sequenced contigs. The average IGD was 96 $\pm$ 128 kb and the median 43 kb. The IGD distribution pattern (Figure 2) indicated that very few genes are separated by a distance of 75 to 100 kb, a size that would correspond to an even distribution of genes along the chromosome (Figure 2). By contrast, short and large IGDs were overrepresented with 40% smaller than 20 kb and 24% larger than 150 kb. Based on these results, we propose to choose the N50 value as a threshold to define gene islands in wheat as clusters of at least two genes separated by <43 kb. Using this threshold, 75% (132) of the annotated genes/pseudogenes belonged to 42 gene islands, whereas 23% (41) could be considered as isolated genes. Thus, these results suggest that a majority of wheat genes are clustered into numerous islands of very small size that contain less than four genes (3.2 $\pm$ 1.6 genes ranging from 2 to 10) on average.

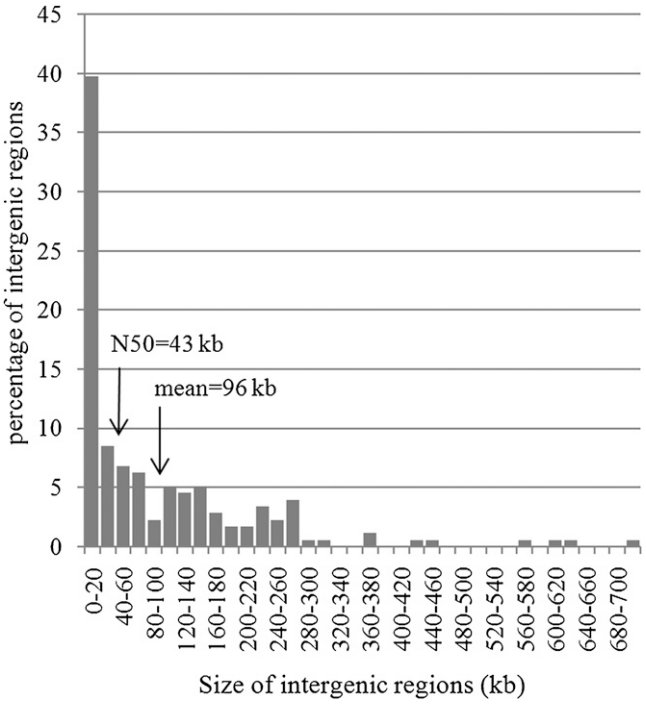


Figure 2. Distribution of the Size of the 176 IGDs.

The *x* axis displays the different size intervals for the IGDs, while the *y* axis presents the percentage of intergenic regions found for each size interval. The mean (96 ± 128 kb) and median (42.5 kb) values are indicated by arrows.

Assessing the 3B Gene Catalog Set

The average gene density of 1 gene/104 kb leads to an estimate of 9560 genes on chromosome 3B, 60% more genes than previously estimated by BAC end sequencing (6000 genes in Paux et al. 2006). Such a high number of genes could be due partly to an overrepresentation of distal contigs with 69% of the sample sequences having two times the gene density of that observed in proximal regions. Considering that distal regions represent half of the chromosome (500 Mb), a weighted value of ~8400 genes on 3B can be estimated (5700 telomeric and 2700 centromeric) for a total of 50,000 per diploid genome.

To get another estimate for the gene number on this chromosome, we used the 1973 Mb of sequences obtained by Solexa/Illumina sequencing of sorted chromosome 3B. First, the 55 million reads were mapped against the 510,160 bp of sequence that corresponds to the 199 low-copy genic regions identified during the annotation of the 18 Mb of contig sequences. The results revealed an average coverage of 1.2x per gene with large variations ranging from 0 to 3.7x, whereas ~25% of the genic regions were covered <0.2 times. The N50 was 0.8x, indicating that despite a theoretical coverage of 2x, only half of the 3B chromosome sequence is covered more than 0.8 times by the Solexa sequence reads. The low GC content (42.8%) of the Solexa sequences compared with the average value (46%) observed in wheat BAC sequences may indicate a preferential amplification of the repeated fraction during the preparation of sorted 3B DNA before sequencing. This would result in an underrepresentation of the genic fraction and the observed discrepancy. The 0.8x value was applied in a second analysis whereby the 1973 Mb of Solexa reads were aligned against the 40,349 sequences present in the wheat unigenes database (NCBI build#55). A total of 2748 unigenes were identified with a coverage of at least 0.8x. Taking into account the N50 value (i.e., 5496 unigenes in total) and the fact that ~24% of the annotated genic regions from the Mb-sized contigs were not found in the wheat unigene set, a total of 7230 genic regions was estimated for chromosome 3B by this whole-chromosome shotgun approach. As it is not possible to distinguish complete genes from gene fragments in the short reads data set, the gene number estimate can be refined by taking into account the proportion of gene fragments observed during the Mb-sized contig annotation (12.1% of the genic regions). This results in an estimate of ~6360 genes for chromosome 3B and, therefore, ~40,000 for the B genome.

Composition and Distribution of the TEs

To obtain novel information on the relative distribution and evolution patterns of TEs along wheat chromosomes and provide a foundation for the future annotation of the wheat genome, particular attention was given to the identification and annotation of TEs. In total, 3222 complete or truncated elements were identified, including 126 known and 818 new families. TEs represented 81.4% of the genomic sample sequences (Table 2; see Supplemental Figure 4 online) and ranged from 46 bp in size for a MITE (a nonautonomous transposon Mariner *Athos*) to 31,157 bp for a *Jorge* CACTA transposon. To correct potential bias in the representation of the sequences, the most gene-rich regions (*ctg0954* and *ctg0011*) and the most repetitive centromeric region (*100L17*) were removed from the statistical analyses. The corrected value indicated 88.2% of TEs, a level higher than the 76.3% estimated previously from the analysis of 11 Mb of BES from the 3B chromosome library (Paux et al., 2006). Class 1 retrotransposons and class 2 DNA transposons accounted for 65.9 and 14.5% of the sequences, respectively (Table 2). This ratio is in between rice, in which class 2 outnumber class 1 elements (International Rice Genome Sequencing Project, 2005), and maize where class 2 represents 8.6% of the genome (Schnable et al., 2009). Interestingly, while the class 1 TE proportion was similar to the 68.7% observed in BESs (Paux et al., 2006), class 2 elements were 3 times more abundant than in the previous estimate (5.6%). The difference was mostly due to the CACTA transposons that represented 13.6% of the 13 contigs sequences (Table 2) versus 4.9% of the BESs (Paux et al., 2006). The detection of CACTA transposons is important not only for understanding genome composition but also because of their role in capturing and shuffling exons (Wicker et al., 2003) as shown for the Pack-MULES and helitrons in rice and maize (Jiang et al., 2004; Morgante et al., 2005). Here, we identified four CACTAs containing three different host gene fragments and one gene likely to be complete. Thus, extrapolating this value to the whole chromosome, CACTAs would be involved potentially in capturing ~200 gene fragments on chromosome 3B, the same

Table 2. Classification of the 3222 TEs Annotated in the Whole 18.2 Mb of Sequence Produced from the 13-Mb Contigs of Chromosome 3B

Type of TE	No. Copies	% of TE Copies	No. Bases	% of TE Fraction	% of Genome Fraction	
LTR retrotransposons						
Gypsy	1,219	37.83	7,939,750	53.23	43.59	64.10
Complete	*688*	*56.44*	*6,264,453*	*78.90*	*34.40*	
Solo LTR	*91*	*7.47*	*175,245*	*2.21*	*0.96*	
Truncated	*367*	*30.11*	*1,195,157*	*15.05*	*6.56*	
Unknown	*73*	*5.99*	*304,895*	*3.84*	*1.67*	
Copia	504	15.64	3,036,081	20.35	16.67	
Complete	*264*	*52.38*	*2,374,144*	*78.20*	*13.04*	
Solo LTR	*35*	*6.94*	*65,294*	*2.15*	*0.36*	
Truncated	*160*	*31.75*	*435,234*	*14.34*	*2.39*	
Unknown	*45*	*8.93*	*161,409*	*5.32*	*0.89*	
Unknown	127	3.94	698,013	4.68	3.83	
Complete	*75*	*59.06*	*583,189*	*83.55*	*3.20*	
Solo LTR	*20*	*15.75*	*33,150*	*4.75*	*0.18*	
Truncated	*29*	*22.83*	*73,461*	*10.52*	*0.40*	
Unknown	*3*	*2.36*	*8,213*	*1.18*	*0.05*	
Non-LTR retrotransposons						
LINE	170	5.28	317,598	2.13	1.74	1.75
SINE	3	0.09	716	0.00	0.00	
Transposons						
CACTA	332	10.30	2,479,618	16.62	13.61	14.53
Complete	*137*	*41.2*	*1,521,871*	*61.38*	*8.36*	
Truncated	*156*	*46.99*	*639,258*	*25.78*	*3.51*	
Unknown	*39*	*11.75*	*318,489*	*12.84*	*1.75*	
Harbinger	35	1.09	26,479	0.18	0.15	
Mariner	294	9.12	38,550	0.26	0.21	
Mutator	74	2.30	43,724	0.29	0.24	
Hat	7	0.22	9,444	0.06	0.05	
Others	193	5.99	49,180	0.33	0.27	
Helitrons						
Helitron	5	0.16	13,562	0.09	0.07	0.07
Unclassified	259	8.04	264,184	1.77	1.45	1.45
Total	3,222		14,916,899		81.90	

The distribution of complete, solo LTR, truncated, and unknown (i.e., undefined or partially sequenced) is indicated for the LTR retrotransposons and the CACTA transposons. For these later subcategories, the proportions of copy numbers and TE fraction are expressed as a percentage within each superfamily.

number reported recently for the *Sorghum bicolor* genome (Paterson et al., 2009). In addition, five helitron-derived elements were found in contigs *ctg0011* and *ctg0954*, but none contained host gene fragments.

Similarity searches against the TREP10 databank (http://wheat.pw.usda.gov/ITMI/Repeats/) and clustering of newly discovered elements enabled the classification of the 3222 TEs into 944 different families containing 1 to 226 members. Despite this wide diversity, only eight families (*Fatima*, *Jorge*, *Angela*, *Laura*, *Sabrina*, *WIS*, *Wilma*, and *Nusif*) account for >50% of the TE fraction (see Supplemental Figure 5A online). Interestingly, 61% of the repeat families showed a very low MDR_{N90} (<10), confirming that the wide majority of TEs is weakly repeated. The elements found in the 818 new families accounted for 13.9% of the total TE fraction. They will greatly enrich the TE databanks and improve future high throughput TE annotations of the wheat genome using automated pipelines.

To investigate evolutionary forces that shaped the wheat genome, we determined precisely the proportions of complete (with two complete TSDs and/or LTRs/TIRs) versus truncated elements. A total of 1027 (59%) LTR retrotransposons and 137 (47%) CACTAs were identified as complete with quite similar ratios between the different sequenced regions (SD =11 and 14%, respectively; Table 2). This result indicates that most of the TEs are still complete in the wheat genome but are highly nested and that target site duplications of an ancient element may be found far apart from each other (e.g., >200 kb away in some cases). When individual BACs were considered within each contig, the proportion of complete LTR retrotransposons and CACTA transposons decreased significantly because in 18% of the cases (up to 50% for small BACs), the rest of the sequence was not present on the same BAC.

TE content, both in terms of amount and type, was highly variable between the different chromosomal locations (see Supplemental Tables 2 and 3 online). The two distal *ctg0011* and *ctg0954* show a very low TE content (59.2%) compared with more proximal ones (*100L17* and *ctg1030*), which contain close to 100% of TEs without any gene. The composition of

the centromeric sequence was very different from the other regions. It consisted exclusively of three types of LTR retrotransposons: *Cereba*, *Quinta*, and a newly identified Copia (*Unnamedfam6*) repeated in 27, 18, and 4 highly nested copies, respectively. The *Quinta* and *Cereba* elements were also found in other proximal contigs (see Supplemental Table 2 online) but at a lower density and with much older insertion dates (1.6 versus 0.6 million years) than those identified at the centromere. Indeed, the youngest average insertion time (0.490 million years ago) among all LTR retrotransposon activities was found for the centromeric BAC elements. These results suggest that the current wheat centromeres have been shaped by the recent reactivation of TEs ancestrally present in the proximal regions. Alternatively, the recent insertion time estimates may originate from a reduced substitution rate around the centromere compared with the other regions. It may also be due to an increased rate of sequence conversion related to a high level of unequal homologous recombination between LTRs as observed previously at rice centromeres (Ma and Bennetzen, 2006).

Other TE families exhibited contrasted distribution patterns: *Barbara* retrotransposons were found mainly in the proximal regions, whereas some CACTAs (*Caspar*, *Clifford*, and *Boris*) were repeated preferentially in the distal parts of the chromosome. Finally, *Laura* retrotransposons were found in significantly higher density in the geneless contig *ctg1030* than in the other contigs (24.0 versus 4.7%). This may reflect preferential insertion of this element in a region (bin 3BS1-0.33-0.55) that contains heterochromatin (Gill et al., 1991), thereby suggesting a potential *Laura* signature for heterochromatic DNA.

Different Waves of TE Amplification Have Shaped the Wheat Genome

LTR nucleotide divergence indicated that 90% of the LTR retroelements transposed <3 million years ago with the oldest element (a Gypsy *Lisa*) inserted 7.5 million years ago and the most recent (two Gypsy *Quinta* elements with 100% identical LTRs) inserted <40,000 years ago (Figure 3A). The burst of amplification peaks at 1.4 million years ago (i.e., before the allopolyploidization event 0.5 million years ago [Huang et al., 2002; Dvorak et al., 2006] at the origin of the tetraploid ancestor *Triticum turgidum*). Furthermore, each family of LTR retrotransposons exhibited a specific pattern of activity (Figure 3B). Most of the highly repeated families showed a single narrow peak (e.g., *Laura*, *Cereba*, and *Fatima*), suggesting that amplification occurred mainly during a short period of time (<1 million years) possibly followed by silencing. By contrast, some families, such as *WIS*, seem to have been amplified over a long period of time (over 2.5 million years). The TE transposition time was also specific to each family with some showing recent transposition activity (e.g., *Cereba*, *Quinta*, and *Laura*; <1 million years ago), while others have been inactive since ~2 million years ago (e.g., *Derami*, *Nusif*, and *WHAM*), indicating that transposition bursts did not involve all TE families at once and that the wheat genome was shaped by a succession of amplification waves involving different families. Moreover, the results show that, except for the centromeric regions, no significant LTR retrotransposon activity has affected the wheat genome in the past 0.5 million years (N10, Figure 3A).

Homogeneous Expansion and Accelerated Evolution of the Wheat Genome

Wheat chromosome 3B is highly collinear to rice chromosome 1 (Sorrells et al., 2003; Salse et al., 2008). Alignment of the 13 contig sequences revealed that they are mostly collinear between the two genomes (Figure 4). We estimated the relative physical distances between the contigs by their position within the deletion bins and compared it with the relative distance of their orthologous regions on rice chromosome 1. This did not reveal any significant difference along the two chromosomes (Figure 4). Since genes generally are evenly distributed along rice chromosome 1, our findings suggest that the proximal cytogenetic bins of wheat chromosome 3B contain as many orthologous genes as the distal ones and that the distal and proximal regions have been expanded to the same extent in wheat. Furthermore, at the scale of the BAC contigs, the size ratio of wheat and rice orthologous regions (14x; 13.1 Mb in wheat versus 0.91 Mb in rice) corresponds to the ratio of size between the wheat B and rice genomes (15x). All regions have expanded at a similar level (14 ± 5x) whatever their proximal or distal location. Thus, despite different burst times and contrasted patterns of insertion, globally, the amplification of TEs occurred at the same intensity across the wheat chromosomes, resulting in an apparently homogeneous expansion of the genome. To confirm this result at the sequence level, we compared the distances between all available adjacent pairs of orthologous genes in wheat and rice. For the 74 analyzed IGDs, we found that only 30% (22/74) had a ratio between 5 and 30x, indicating an expansion level around the average value. By contrast, 40% (30/74) showed little or no expansion (ratio <5), thereby contributing to the formation of gene islands, while 30% (22/74) have massively expanded (ratio >30). We did not observe any differences for this pattern between the distal and proximal regions. Thus, we conclude that the average expansion factor of 14x observed between the wheat and rice genomes does not correspond to a globally homogeneous expansion but encompasses great local variations that can be of several orders of magnitude.

To further investigate the evolution patterns along wheat chromosome 3B, we compared systematically the 175 wheat predicted genes and pseudogenes with the rice and *Brachypodium distachyon* genome sequences. Only half (52%, 91 genes including six pseudogenes) were strictly orthologous with genes of rice chromosome 1 or of *B. distachyon* chromosome 2 (Figure 5A). Their products shared 77% ± 12% and 82% ± 12% amino acid identity (see Supplemental Table 1 online), respectively. Among the 91 orthologous genes, 88 were conserved and syntenic between the three genomes, two were common with rice but absent in *B. distachyon*, and one was conserved with *B. distachyon* while absent in rice (Figure 5A). These 91 genes represent an ancestral *Poaceae* genic backbone that exhibits a constant gene density of 1 gene per 200 kb (Figure 1) across the proximal and distal contigs. This is reflected in the conserved homogeneous distribution of genes observed through the alignment of the wheat 3B and rice 1 chromosomes (Figure 4).

In addition to this ancestral backbone, wheat contigs carry an unexpectedly high proportion (48% of the gene content; Figure 5) of non collinear genes. While 21 out of the 84 noncollinear genes

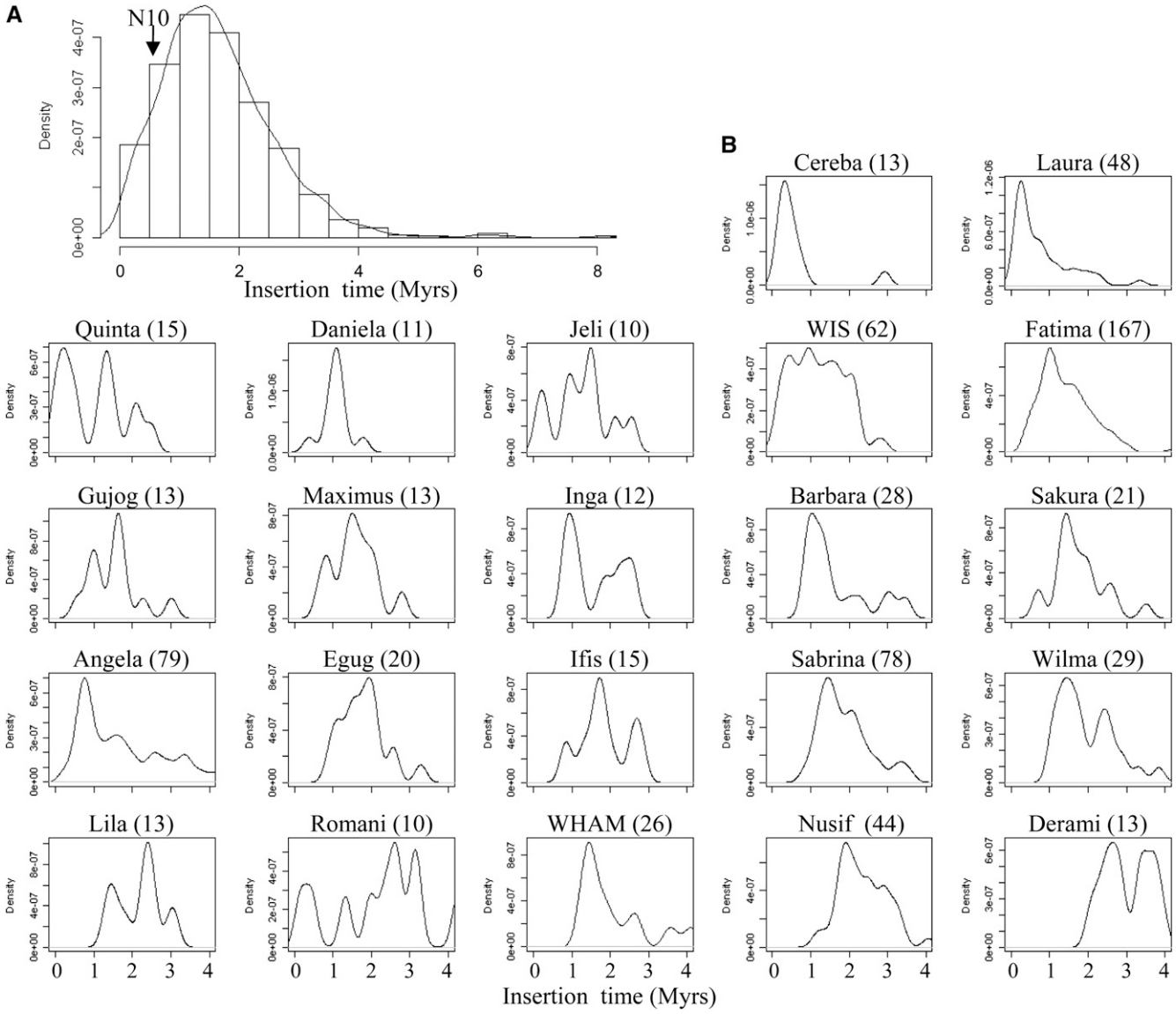


Figure 3. Ages of LTR-Retrotransposon Insertions.

Distribution of the age of the insertion of 880 complete LTR retrotransposons **(A)** and the 13 most abundant LTR retrotransposon families **(B)**. The insertion time was calculated based on the LTR sequence divergence using a substitution rate of 1.3×10^{-8} substitutions/site/year (Ma and Bennetzen, 2004). For each family, the number of complete copies used to calculate the insertion pattern is indicated in brackets. The cutoff value of 0.5 million years (Myrs) distinguishing the 10% of youngest elements (N10) is indicated with an arrow.

are likely pseudogenes, the majority (75%) shows complete open reading frames (ORFs) and significant similarities over >70% of their length with a homologous gene product in rice and *B. distachyon*, indicating that they do not correspond to wrongly annotated repetitive elements. Noncollinear genes were found in all contigs but were more abundant in the subtelomeric contigs (Figure 1). *Ctg0954* and *ctg0011* showed the highest degree of noncollinearity with 62 and 87% (33/53 and 27/31), respectively, of their coding fraction not found in orthologous regions of the other genomes. By contrast, the whole set of 13 genes carried by *ctg0079* was collinear in *B. distachyon* (only one gene missing in rice). Thus, this result strongly suggests that the twofold increase in gene density observed in the subtelomeric regions is due mainly to the presence of a high amount of noncollinear genes. Interestingly, nonsyntenic genes were not found in clusters; rather, they were interspersed along the ancestral gene backbone, thus disrupting collinearity with rice and *B. distachyon* at many locations (Figure 5B). One of the most striking examples is contig *ctg0954* in which 29 wheat locus-specific genes are interspersed along the ancestral conserved backbone composed of 18 orthologous genes (see Supplemental Figure 6 online). Comparison with the orthologous region that was

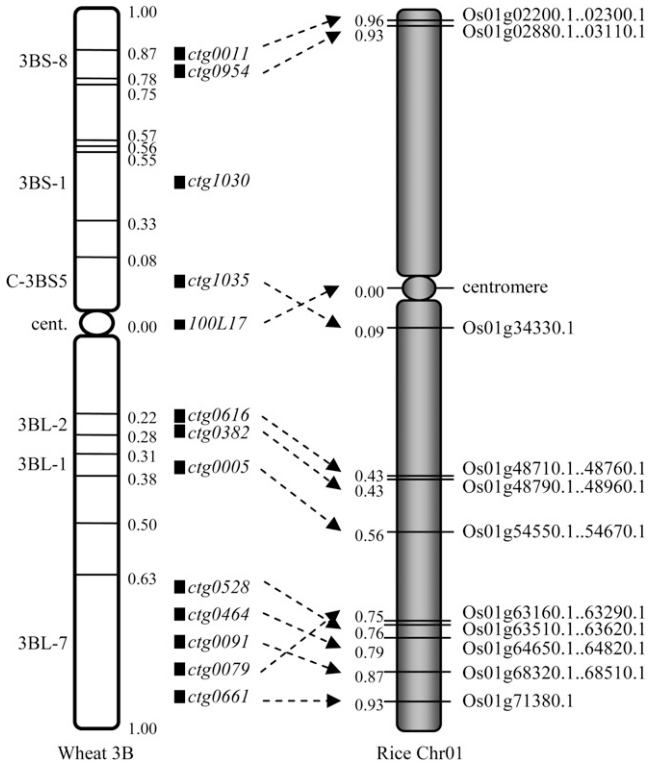


Figure 4. Chromosomal Location of the 13 Sequenced Contigs from the Wheat 3B Chromosome and Their Orthologous Regions on Chromosome 1 of Rice.

The boundaries of 16 deletion bins are indicated by horizontal lines across the 3B chromosome, and their distances from the centromere are expressed as a fraction of length of chromosomal arms. In rice chromosome 1, the first and last genes carried by the orthologous regions are indicated. Their distances from the centromere were calculated using a centromere position at 16.8 Mb (http://rice.plantbiology.msu.edu/pseudomolecules/centromere.shtml). The relative order of the contigs in bin 3BL7-0.63-1.00 was determined by genetic mapping (data not shown).

partially sequenced at the *Rph7* locus in barley (*Hordeum vulgare*; Brunner et al., 2003) showed that noncollinear wheat genes are conserved in barley. In addition, comparison of the 84 noncollinear wheat genes with the barley mapped ESTs (http://www.harvest-web.org; assembly #35) revealed that half of them are on chromosome 3H and are likely syntenic with wheat. Thus, our results suggest that the high level of gene rearrangements observed in wheat compared with rice and *B. distachyon* is a common feature of the *Triticeae*.

The noncollinear wheat genes shared similarities with genes located on several other chromosomes in rice and *B. distachyon*, suggesting that they originate mainly from independent events of translocation or interchromosomal duplication in the wheat genome. Moreover, for these genes, the best BLAST hits were collinear in the rice and *B. distachyon* genomes in 80% of the cases revealing ancestral loci that have been duplicated and/or translocated to new chromosomal locations in wheat specifically. Interestingly, a significant correlation (r^2 = 0.744) was observed between the level of collinearity disruption and the level of TE activity (see Supplemental Figure 7 online) with the most rearranged contigs (*ctg0011*, *ctg0954*, and *ctg0661*) showing the most recent (1.0 to 1.4 million years on average) TE activity, while the most conserved regions (*ctg0528*, *ctg0079*, and *ctg0005*) contained the oldest TEs (1.9 to 2.1 million years on average). Thus, together, these results suggest that interchromosomal duplications mediated by transposable elements had a major impact on increasing the number of genes and in relocating genes at nonsyntenic regions in the wheat and most probably the ancestral *Triticeae* genome.

We also examined the ratio of genes found in the rice and *B. distachyon* orthologous regions but not present on chromosome 3B: only nine nonconserved genes were found in rice and 11 in *B. distachyon* (Figure 5A). All of the nonconserved genes, except the pair Os01g48810/Bradi2g46610, were not orthologous between rice and *B. distachyon*, indicating that their absence in wheat does not reflect a deletion from an ancestral gene in the *Triticum* lineage but rather a gene insertion specifically in the rice or *B. distachyon* lineages (Figure 5B). The analysis of the complete genomes of rice and *B. distachyon* showed that these locus-specific genes are duplicated within their respective genomes, suggesting a predominant role for interchromosomal duplications in gene movement in these genomes. Further

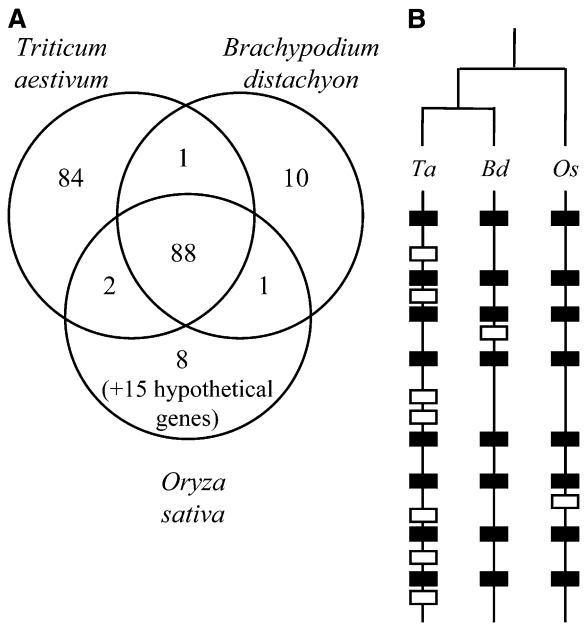


Figure 5. Level of Synteny between Wheat, Rice and *B. distachyon* Genomes.

(A) Venn diagram of the syntenic and nonsyntenic genes between wheat, rice, and *B. distachyon*. Fifteen additional hypothetical rice genes that do not share any similarity within the sequence databanks and represent putative prediction errors are mentioned on the diagram. All other nonsyntenic genes identified (84 in wheat, 10 in *B. distachyon*, and 8 in rice) have homologs in the compared species.

(B) Schematic representation of orthologous chromosomes displaying orthologous (black) and nonsyntenic (white) genes in wheat (Ta), *B. distachyon* (Bd), and rice (Os).

analyses are underway to determine whether this is the same in wheat, in particular whether homologs of the noncollinear genes are also found and in which proportion on other wheat chromosomes.

Tandem duplicated genes represented 33% of the gene content in our data set with 26 groups of duplicated genes composed of two to seven copies. Tandem gene duplications occurred primarily in the telomeric regions with 64 of the 66 duplicated genes found in six distal contigs. Only three pairs of duplicated genes were found duplicated in rice and *B. distachyon* as well, revealing that they were ancestrally duplicated and that several copies are maintained by selection following ancestral sub- or neofunctionalization. The 23 remaining tandem duplicated groups are found in a single copy at syntenic positions in rice and *B. distachyon*, indicating that most of the gene duplications occurred recently in the *Triticeae* evolution. These results demonstrate that, in addition to interchromosomal duplications, tandem gene duplication in distal regions is a major force driving gene number increases in wheat with a strong potential for the creation of new functions. Most (18/26) of the duplicated groups have at least one gene copy that is a pseudogene, revealing that, in most of cases, gene duplication does not increase fitness sufficiently enough to result in both copies being maintained by selection. Interestingly, among the 51 pseudogenes and gene fragments identified, the vast majority (88%; 45/51) were nonsyntenic, indicating that pseudogenization concerned mainly the additional noncollinear genes.

In conclusion, considering that the ancestral gene content remains highly conserved and that half of the coding potential consists of genes inserted recently, our data support the idea that the wheat genome contains more genes than the rice and *B. distachyon* genomes.

DISCUSSION

Mb Level Sequencing Provides Insights into the Wheat Genome Composition

Following the establishment of the first physical map of a wheat chromosome (Paux et al., 2008), we produced and analyzed 18 Mb of contiguous sequences and 2 Gb of whole chromosome survey short reads from chromosome 3B of hexaploid wheat. Together with the 11 Mb of BES obtained previously on the same chromosome (Paux et al., 2006), this unique sample represents a large and diverse sampling of wheat sequences compared with other sequences produced thus far, which include only a few contigs larger than 300 kb. Comparison of the features from the three different sequence data sets indicated putative bias induced by different approaches in assessing genome composition. Evidence for bias was determined first with GC content estimates. While contigs displayed a similar GC content (46%) as the *Triticum/Aegilops* BAC sequences present in the public databases, BES and Solexa sequences showed significantly lower values, suggesting a potential bias for AT-rich regions in these samples. This was particularly true for the Solexa sequencing sample in which amplification of sorted chromosomal DNA prior to sequencing seems to be less efficient for the genic regions (GC-rich) than for the TE fraction (AT-rich). For the BESs, the reduced GC% is likely related to an intrinsic feature of the wheat genome in which *Hin*dIII restriction sites (used to produce the wheat BAC libraries) are overrepresented in eight of the most repeated TE families that display a low GC content (F. Choulet, unpublished data).

Another discrepancy was found in the estimation of the percentage of CACTA elements. In this study and in a recent analysis of 10 single wheat BAC clones from chromosome 3B (Charles et al., 2008), 3 times more CACTA elements were identified than in the BES survey of the same chromosome (Paux et al., 2006). CACTAs are particularly prone to underestimation because their sequences are highly variable (Wicker et al., 2008), and it is more difficult to identify them by similarity from short sequences than from long stretches of DNA. Furthermore, having the ability to analyze large contiguous sequences enabled us to find that even if they are highly nested, the majority of CACTAs and LTR retrotransposons are still complete in the wheat genome. This feature was not observed in previous studies based on individual BAC sequencing (Charles et al., 2008) because the nested clusters generally span distances that are larger than a BAC. Combined with the fact that most TEs transposed more than 0.5 million years ago (i.e., before the two polyploidization events), these results support the idea that deletion forces have been low in comparison to those of amplification during evolution, thereby contributing to the large size of the wheat genome (Wicker and Keller, 2007).

Finally, significant variations between the different samples were found when estimating the number of genes. Depending on the sequence sample, between 6000 and 8400 genes (+1000 gene fragments) were predicted for chromosome 3B; this would translate to 36,000 to 50,000 for the B genome of hexaploid wheat. This is slightly higher than the 32,000 to 40,000 annotated genes obtained from the rice, sorghum, and maize whole-genome sequences (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009). The lower estimate for wheat is in the range of the number of unigenes (40,349) deduced from the EST collection, which by definition cannot be complete (NCBI build#55). On the other hand, the upper estimate is still lower than the 98,000 genes per diploid genomes estimated through random sequencing of wheat methyl-filtrated libraries (Rabinowicz et al., 2005; Paux et al., 2006). These discrepancies can have different origins. The lower estimate was obtained from BESs and is most likely an underestimation. Indeed, we found evidence that due to the increased frequency of *Hin*dIII sites in some TE families, BESs obtained from the *Hin*dIII 3B-specific BAC library (Safar et al., 2004) originate preferentially from the repeated fraction of the chromosome, thereby leading to an underestimation of the gene content. We also believe that the whole 3B survey using Solexa reads (7230 genes) slightly underestimates the gene number because the coverage is highly uneven when amplification is used prior to sequencing and genic regions tend to be less represented. Thus, assessing the gene content by this approach is rather unreliable. Finally, we believe that the precise annotation of the 13 Mb-sized contigs provides the best estimate at this time because complete gene structures with ORF, pseudogenes, and gene fragments as well as different members of tandemly

repeated gene families could be distinguished and taken into account in the gene number calculation. This level of resolution cannot be achieved with short sequence surveys that collapse paralogs and cannot distinguish gene fragments from complete genes. Similar conclusions were reached for maize, where the characterization of the gene and repeat spaces based on genome survey sequences (Messing et al., 2004) was further refined by sequencing large BAC or contig sequences (Haberer et al., 2005; Kronmiller and Wise, 2009).

Thus, even if our sample may not be completely representative of the whole genome, we conclude that, in addition to faster and more random approaches, megabase-level sequencing provides essential knowledge of the genome composition and organization for the delineation of the best strategy for sequencing the entire genome. Individual chromosome sequencing efforts that are currently underway at the international level (see www.wheatgenome.org) will help to confirm and refine our estimates for the whole genome in the near future.

Genes Are Mainly Clustered into Small Islands Spread Out along the 3B Chromosome

Previous work based on EST mapping into deletion bins concluded that 94% of wheat genes are clustered into gene-rich regions spanning 29% of the chromosomes (Erayman et al., 2004). In addition, since crossing-over frequency is related to gene density (Akhunov et al., 2003b), it was suggested that centromeric regions, where recombination is largely suppressed in wheat, could be devoid of genes. Here, Mb-sized contig sequences and MTP macroarray hybridizations show that regions larger than 800 kb without genes are rare on chromosome 3B, even in the proximal regions. This refines the findings of Devos et al. (2005) and Charles et al. (2008) who also observed the presence of one or two genes per BAC in a majority of clones randomly chosen from whole-genome and chromosome 3B-specific libraries. This finding has great implications for the future sequencing of the wheat genome since accessing the entire wheat gene space will require sequencing at least 90% of the BAC contigs obtained for the minimal tiling paths of the different wheat chromosomes. Therefore, cost-efficient strategies need to be established to ensure complete sequencing of these contigs. Spatial gene distribution and sequencing strategy were refined on the same model for the maize genome after the first large sequences were analyzed (Haberer et al., 2005; Liu et al., 2007; Kronmiller and Wise, 2009) and revealed a much more homogeneous distribution than proposed previously (Carels et al., 1995; Barakat et al., 1997).

The possibility of examining large contiguous sequences from contrasted regions also enabled us to better define the gene island concept in wheat. Gene islands reflect inhomogeneous expansion of the genome and are not found in compact genome species such as rice, *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), and *Brachypodium* (Huo et al., 2009). They are, however, common features of large and repetitive genomes, such as the 2.5 Gb maize genome. Interestingly, wheat and maize share similar genome structures despite divergent evolutionary histories (Salse et al., 2009). In maize, the gene distribution is quite similar to wheat with small islands (one to four genes/island) separated by blocks of repetitive elements and only 22% of the annotated BACs without genes (compared with 27% in our sample; Kronmiller and Wise, 2009). This suggests that common evolutionary forces govern the dynamics of large plant genomes. Gene islands may originate from two different evolutionary scenarios: (1) selection against the separation of genes by TE insertions that would be deleterious for gene expression or regulation and (2) homogeneous expansion combined with preferential deletions in gene-rich regions. The latter was proposed as the explanation for increased gene density in the distal regions of the sorghum chromosomes (Paterson et al., 2009). In maize, old TEs were found more frequently in BACs without genes than in gene-rich BACs, suggesting that the elimination rate of LTR retroelement is higher in gene-rich regions (Liu et al., 2007). In fact, since old elements are enlarged by subsequent insertions of younger TEs, the probability of finding old elements is lower in gene-rich than in gene-free BACs simply because the intergenic spaces studied are smaller. This is an intrinsic limit of sequencing isolated BAC clones as opposed to large stretches of contiguous sequence. In our sample, old LTR retroelements were found at every locus (except the centromeric BAC). TE deletion also can be monitored partly through the identification of solo-LTRs (Devos et al., 2002). Here, they were found mainly in the large blocks of TEs rather than in the gene islands, suggesting that TE removal is not the main driver for the formation of gene islands in wheat (Tian et al., 2009). The apparently homogeneous expansion of the wheat genome resulted from a combination of massive expansion of some regions that accumulated TEs and may have served as a nucleation point for additional TE insertions with selection against TE insertion maintaining a majority of genes in close vicinity from each other. To investigate further the dynamics of gene islands in wheat, we are developing a transcriptional map of the wheat 3B chromosome, which will permit us to determine whether genes located in gene islands are preferentially expressed or coregulated when compared with more isolated genes.

Additional Noncollinear Genes Are Interspersed within a Very Conserved Ancestral Grass Gene Backbone

Our results indicate a gene density that is two times higher in the distal (1 gene/87 kb) than in the proximal contigs (1 gene/184 kb). At the same time, we observed that the relative distances of the genes to the centromeres were similar on the wheat 3B and rice 1 chromosomes, indicating that the distribution of genes belonging to the ancestral *Poaceae* backbone is homogeneous in the two species. In addition, the size ratio of wheat and rice orthologous regions (14x) revealed that despite different burst times and nonrandom patterns of insertion, the amplification of TEs occurred at the same intensity across the wheat chromosomes, resulting in a homogeneous expansion of the proximal and distal regions. Thus, differential efficiency of transposition cannot explain the increased gene density toward the telomeres. By contrast, our findings suggest that TE-mediated interchromosomal and tandem gene duplications are primarily responsible for the higher gene density and higher disruption of collinearity in the telomeric regions of wheat chromosome 3B. Comparative sequence analyses revealed that 99% of the genes

that were orthologous between rice and *B. distachyon* were also found at syntenic locations on chromosome 3B, revealing that no rearrangement of the ancestral gene order has occurred over the whole 18 Mb of wheat contigs. However, a high proportion of nonsyntenic genes was observed in the distal regions, especially on 3BS. All nonsyntenic genes shared significant similarity with genes in rice or *B. distachyon* genomes located on different chromosomes, indicating that they may originate from interchromosomal duplications of DNA fragments carrying complete or partial genes as previously suggested by Li and Gill (2002) and more recently by Akhunov et al. (2007). The fact that the nonsyntenic genes are not clustered but rather interspersed along the ancestral backbone also supports the idea of multiple events of long-distance duplications. This is in perfect agreement with previous results based on EST hybridizations that showed that 25% of gene loci are duplicated (either interchromosomal or intrachromosomal) in wheat, especially in the distal regions (Akhunov et al., 2003b). In addition, one-third on the genes were found tandemly duplicated within the distal contigs with a majority of nonsyntenic genes, suggesting that, in addition to translocations, tandem gene duplications played a significant role in increasing the number of genes at the telomeric ends of the chromosomes during wheat evolution.

High duplication and translocation activities that occurred in the wheat and most likely barley lineages resulted in an increased level of synteny perturbations, which was suggested previously by EST mapping for the A and D genomes (Akhunov et al., 2003a) and in the distal bin of 4BL (See et al., 2006). Here, we observed that the rice and *B. distachyon* genomes are much more similar to each other, in terms of gene content, than they are to wheat. This confirms at the sequence level the accelerated evolution in the *Triticeae* that was suggested very recently through comparison of an EST genetic map of *A. tauschii* with the rice and sorghum genomes (Luo et al., 2009). Thus, although *Brachypodium* is phylogenetically closer to wheat than rice (Griffiths et al., 2006) and sequence homology is higher, the lineage-specific rearrangements that occurred in wheat have disrupted the synteny as much with *Brachypodium* as with rice. Therefore, for structural, synteny-based genomic analyses, *Brachypodium* will not provide significantly better support than rice for estimating the gene order and content of the wheat chromosomes.

Evidence for TE-mediated interchromosomal duplications was obtained from two types of results. First, there was a clear correlation between the levels of nonsyntenic genes and TE activity found in the telomeric regions. Second, we identified four genes that were included in CACTA transposons. CACTAs are rare in the rice (0.3%) and maize genomes (3%) where gene capture has been driven massively by Pack-Mules and helitrons (Jiang et al., 2004; Morgante et al., 2005), although CACTAs carrying gene fragments were already observed in maize (Li et al., 2009). By contrast, they were estimated to account for 14% of the wheat B genome, suggesting that, as was shown recently for sorghum (Paterson et al., 2009), CACTA-mediated gene capture may be one of the main mechanisms for gene amplification and mobilization in wheat (Akhunov et al., 2007).

Following a normal evolutionary pattern, we would expect that the 91 genes conserved at regions syntenic between wheat, rice, and *B. distachyon* were present in each of the A, B, and D diploid ancestral homeologous genomes. Interestingly, only a very small fraction (7%) of these genes indicated pseudogenization, suggesting that despite different episodes of polyploidization and segmental duplications, the wheat genes were not affected by major structural rearrangements as previously observed by sequencing homeologous loci in wheat (Chalupska et al., 2008). Thus, in contrast with paleopolyploids such as maize, gene loss has not been that extensive in wheat since its polyploid origins, confirming that the mechanisms of coping with polyploidization have varied significantly for different genomes (Hufton and Panopoulou, 2009). Polyploidization events may have occurred too recently (~500,000 years) to observe massive loss of redundant functions by genetic drift; however, a more probable explanation is that most of the duplicated gene copies have been maintained by selection since they represent more than just redundant gene functions. Indeed, there is accumulating evidence for differential expression of homoeologous genes in wheat (Bottley et al., 2006; Pumphrey et al., 2009), suggesting that subfunctionalization is playing a major role in maintaining the structural integrity of duplicated genes along the wheat chromosomes.

Here, by analyzing a unique set of Mb-sized contiguous sequences and of whole chromosome survey sequences from a wheat chromosome, we gained extensive insights into the composition, organization, and evolution of the wheat genome that will permit us to better define the genome sequencing and annotation strategies. With the development of additional physical maps and the sequencing of the first chromosomes in the near future, we will be able to refine our knowledge and develop even better tools to access and exploit the wheat genome. With one Insertion Site-Based Polymorphism marker every 3.8 kb and one Simple Sequence Repeat every 13 kb (Paux et al., 2010), the wheat genome sequence holds the potential for unlimited marker development and for a paradigm shift in breeding. The preliminary work presented here is paving the way toward enabling this transformative technology for wheat.

METHODS

BAC Screening and Contig Selection

In total, 152 BACs (see Supplemental Table 4 online) from the Minimal Tiling Path of 13 contigs selected from the physical map of chromosome 3B (Paux et al., 2008) were used for complete sequencing. Sequences and annotation are available through the Wheat 3B Physical Map Genome Browser at http://urgi.versailles.inra.fr/cgi-bin/gbrowse/wheat_FPC_pub/. The contigs were chosen to cover different regions of the 3B chromosome. Two contigs (*ctg0011* and *ctg0954*) originated from the subtelomeric deletion bin (3BS8-0.778-0.87) on the short arm of the chromosome and were selected from a 12-centimorgan region carrying disease resistance gene. One contig (*ctg1030*) was located in a bin (3BS1-0.33-0.55) positioned at half of the 3BS chromosome arm. A contig (*ctg1035*) and a single BAC clone (TaaCsp3BFhA_0100L17 referred as *100L17*) identified in pericentromeric and centromeric regions were selected to investigate more specifically the composition of centromeric sequences. On the long arm, two contigs (*ctg0616* and *ctg0382*) originated from a subcentromeric region (bin 3BL2-0.22-0.28), and another one (*ctg0005*) was assigned to a more distal deletion bin (3BL1-0.31-0.38) at 1/3 of the chromosome arm. Finally, five contigs were chosen based on

their EST content after screening the 3B BAC library with 399 ESTs assigned to the most distal deletion bin 3BL7-0.63-1.00 (229 ESTs previously assigned; Qi et al., 2004) and an additional 170 markers identified using synteny with rice chromosome 1. Three contigs (*ctg0464*, *ctg0079*, and *ctg0661*) carried two or more ESTs, one carried one EST (*ctg0528*), and one did not carry any EST (*ctg0091*).

BAC Sequencing and Assembly

Ten of the 13 contigs (*ctg0005*, *ctg0079*, *ctg0091*, *ctg0382*, *ctg0464*, *100L17*, *ctg0528*, *ctg0616*, *ctg0661*, and *ctg1035*) were sequenced at the Centre National de Séquençage (Evry, France). First, 101 BACs were sequenced using Sanger technology with libraries obtained after mechanical shearing of BAC DNA and 5-kb fragment cloning into pcdna 2.1 plasmid vector (Invitrogen). DNA was purified and end-sequenced using dye terminator chemistry on ABI 3730 sequencers (Applied Biosystems) at 12× read coverage. Three other BACs were sequenced using 454-GS-FLX sequencer at 25× read coverage (Roche). Contigs *ctg1030* and *ctg0954* were sequenced by GATC Biotech (Konstanz, Germany) using Sanger technology at 8× read coverage. Finally, *ctg0011* was sequenced using Sanger at 6× read coverage. BAC sequences were assembled using Phred/Phrap/Consed (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998) for Sanger reads and Newbler (Roche) for 454 GS-FLX reads. Finishing sequencing reactions were performed for all clones until all contigs could be ordered and orientated for building a single supercontig for each BAC contig.

Annotation and Sequence Analyses

The TriAnnot pipeline (http://urgi.versailles.inra.fr/projects/TriAnnot/) was used for automatic annotation of genes. CDS predictions were combined with similarity searches using NCBI-BLAST (Altschul et al., 1997) against full-length cDNAs, unigenes, and ESTs from wheat (*Triticum aestivum*), *Triticeae*, and other *Poaceae* and against SwissProt and the rice (*Orzya sativa*) proteome specifically. Matching transcripts were mapped on the genomic DNA using Gmap (Wu and Watanabe, 2005). Predicted CDSs that do not share any significant similarity with any sequence in the databanks were discarded from the annotation. Gene models displaying translational stop codons, frameshift mutations, or small deletions (up to 30% of a complete homolog) within the ORF were considered as pseudogenes. Genes showing similarity over <50% of the length of their best homolog in databanks were considered as gene fragments. tRNA genes were identified using tRNAscan-SE (Lowe and Eddy, 1997).

For TE annotation, RepeatMasker Open (Smit et al., 1996–2004; http://www.repeatmasker.org) was used to find similarities against the TREP databank (http://wheat.pw.usda.gov/ITMI/Repeats/), and the dotter program (Sonnhammer and Durbin, 1995) was used to precisely annotate the exact borders of each TE by identifying long terminal repeats or terminal inverted repeats and the target site duplication. Reconstruction of the nested structures of TEs was manually curated under Artemis (Rutherford et al., 2000). Curated annotations were then inserted into our local relational database. Classification of TE was performed by following the procedure described by Wicker et al. (2007). For elements larger than 1 kb, a strong hit (≥80% identity) with an element from TREP and covering at least 500 bp was considered as threshold to assign a family. For elements smaller than 1 kb, a strong hit (≥80% identity) covering at least 50% of the query length was considered. A clustering of unknown elements was performed using exactly the same criteria to identify members of the same new family. Insertion dates of LTR retrotransposons were estimated by aligning both 5′ and 3′ LTRs using ClustalW (Larkin et al., 2007) and considering a mutation rate of 1.3×10^{-8} substitutions/site/year (SanMiguel et al., 1998; Ma and Bennetzen, 2004). This estimation was performed on 880 complete LTR retrotransposons containing no sequencing gap and for which both LTRs do not differ by >200 bp in size (caused by large insertion/deletion).

Solexa Sequencing of Sorted Chromosome 3B, Computing of an MDR Index, and Mapping Short Reads on Reference Sequences

The 3B chromosomes were sorted by flow cytometry and then amplified using Phi29 polymerase. DNA was then sequenced on a flow cell of Illumina/Solexa Genome Analyzer II using paired-end reads (average fragment size: 626 bp). Sequencing generated 54,808,646 paired-end reads of 36 nucleotides each, thus providing 1,973,111,256 nucleotides corresponding to theoretical 2X coverage of the complete chromosome. The complete sample was used for generating a MDR index using the Tallymer program (Kurtz et al., 2008). The optimal k-mer size was evaluated to 17 nucleotides, thus providing more than 1 billion 17-mers from the complete sample of which 81% are not unique. Tallymer was used to count the occurrence of each 17-mer (and its reverse complement) along the sequenced contigs according to the 3B MDR index. We developed a program (*PlotMdr.pl*) (available on request) that parses Tallymer results to generate MDR plots and compute the MDR_{N90} (i.e., the cutoff for the highest 10% of MDR values).

To estimate the real coverage of Solexa read data set, we selected 596 kb low-copy DNA regions corresponding to the 199 genic regions (introns and exons) annotated in the 13 completely sequenced contigs. To avoid unspecific alignment of 36-mers on repeated DNA, we first masked the regions showing MDR ≥5 (representing 86 kb). The addSolexareads script from the Consed suite was then used to map the 55 million of Solexa reads on the resulting 510 kb of reference sequences. The number of reads that correctly map to the template sequences was calculated to estimate the coverage of the Solexa data set. Then, the same procedure was applied to the 40,349 wheat unigenes (NCBI build#55) to identify and estimate the number of genes carried by chromosome 3B.

Hybridization on MTP-3B Macroarray

The 7440 BACs of the wheat chromosome 3B MTP (Paux et al., 2008) were gridded in duplicate on 22 × 22 nylon membranes (Proteigene) in 5 × 5 spot arrays with the control plate. The spotted membranes were incubated at 37°C on Luria-Bertani agar with 12.5 μg/mL chloramphenicol and were treated as described by Paux et al. (2004). Total RNAs of hexaploid wheat cv Chinese Spring were extracted from 500 mg of five organs (root, leaf, stem, spike, and grain) at two or three developmental stages each as described by Cossegal et al. (2008). DNA was removed using RQ1 RNase-free DNase (Promega). The RNeasy MinElute Cleanup kit (Qiagen) was used for purification, and cDNA synthesis was performed with the SMART PCR cDNA synthesis kit (Clontech) followed by purification with the QIAquick PCR purification kit (Qiagen).

DNA probes were labeled with [α-^{32}P]dCTP (Perkin-Elmer; 50 ng SMART PCR cDNA and 50 pg Desmin cDNA) using the Megaprime DNA labeling system (GE Healthcare). Genomic DNA of wheat cv Chinese Spring was partially digested with RQ1 RNase-free DNase (Promega) at 37°C for 30 s, and the reaction was stopped by adding 2 μL of Stop Solution and incubating at 85°C for 15 min. Resulting DNA fragments (between 200 and 400 bp) were used as blocking DNA. Labeled cDNA probes were mixed with 2.5 μg blocking DNA (50× more than cDNA). The mix was denatured at 100°C for 5 min and incubated at 37°C for 30 min before hybridization. Hybridizations were performed as described by Paux et al. (2004). The wrapped filters were exposed for 1 to 2 d (Imaging Screen K; Bio-Rad). A hybridization step with labeled pIndigoBAC-5 polylinker was performed to validate filters quality and to normalize hybridization data.

Imaging screens were scanned with a Pharos FX Plus molecular imager (Bio-Rad) at 50-μm resolution, and ArrayVision 8.0 imaging software

(Imaging Research) was used for signal detection and quantification. MTM density quantification method was chosen, and background for each grid was calculated based on the empty spots using median density method. The raw hybridization signals were normalized as described by Paux et al. (2004). Normalized intensity of each spot was divided by the normalized intensity of the same spot from the pIndigoBAC-5 polylinker hybridization to correct for bacteria growth bias. The median was calculated with the four growth-corrected normalized intensities of each BAC. Intensities above 4 times and below one-quarter of the median were excluded from the analysis. The corrected median was calculated with the remaining growth-corrected normalized intensities of each BAC. Unpaired Welch's two samples *t* tests were performed to compare the intensities of each BAC to the intensities of the negative controls using R (http://www.r-project.org/). A BAC was considered as positive if its P value was below 0.05 and its corrected median above the maximal corrected median of expressed transposable elements controls.

Accession Numbers

BAC contig sequence data from this article can be found in the EMBL/GenBank data libraries under accession numbers FN564426-37 and FN645450. Solexa sequence data can be found in the EMBL Sequence Read Archive under accession number ERA000182.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Graphical Representation of the Annotation of the 13-Mb Contig Sequences from Wheat Chromosome 3B.

Supplemental Figure 2. Distribution of the Number of Exons per Gene.

Supplemental Figure 3. Distribution of the Gene and MITE Densities along the 3.1-Mb *ctg0954*.

Supplemental Figure 4. Histograms of the Composition in Transposable Elements of the 13 Sequenced Contigs.

Supplemental Figure 5. Proportion and MDR Analysis of the TE Families.

Supplemental Figure 6. Sequence Comparison of *ctg0954* with the Rice Orthologous Region on Chromosome 1.

Supplemental Figure 7. Correlation between the Level of Synteny and the Level of TE Activity.

Supplemental Table 1. List of the Genes, Pseudogenes, and Gene Fragments Identified in the 13 Sequenced Contigs of Wheat Chromosome 3B.

Supplemental Table 2. Proportions of the 10 Most Represented TE Families within the 13 Contigs.

Supplemental Table 3. Proportions of the Known TE Families within Distal and Proximal Contigs.

Supplemental Table 4. List of the 152 Sequenced BAC Clones.

ACKNOWLEDGMENTS

We thank D. Boyer from the Institut National de la Recherche Agronomique, Genetics Diversity and Ecophysiology of Cereals, for excellent technical assistance. We also thank P. Wincker, S. Samain, and V. Barbe from the Centre National de Séquençage (Evry, France) for their work on BAC sequencing. We thank I. Gut, C. Plançon, and Y. Duffourd from the Centre National de Génotypage (Evry, France) for their work on Illumina/Solexa sequencing and J. Doležel and H. Šimková (Institute of Experimental Botany, Olomouc, Czech Republic) for the sorted chromosome 3B DNA. This project was supported by grants from the Commissariat à l'Energie Atomique-Genoscope (AP2008), the Agence Nationale de la Recherche ANR-GPLA06001G SMART in the frame of the national Genoplante program, the Institut National de la Recherche Agronomique (AIP Sequencing "Plant ReSeq"), a Turkish Academy of Sciences - Outstanding Young Scientists Award Young investigator award, and the Swiss National Science Foundation (Grant 105620).

Received January 21, 2010; revised May 26, 2010; accepted June 8, 2010; published June 25, 2010.

REFERENCES

Akhunov, E.D., Akhunova, A.R., and Dvorak, J. (2007). Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. Mol. Biol. Evol. **24:** 539–550.

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

Akhunov E.D., et al. (2003b). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. **13:** 753–763.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

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

Barakat, A., Carels, N., and Bernardi, G. (1997). The distribution of genes in the genomes of Gramineae. Proc. Natl. Acad. Sci. USA **94:** 6857–6861.

Bottley, A., Xia, G.M., and Koebner, R.M. (2006). Homoeologous gene silencing in hexaploid wheat. Plant J. **47:** 897–906.

Brunner, S., Keller, B., and Feuillet, C. (2003). A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the *Rph7* locus. Genetics **164:** 673–683.

Carels, N., Barakat, A., and Bernardi, G. (1995). The gene distribution of the maize genome. Proc. Natl. Acad. Sci. USA **92:** 11057–11060.

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. USA **105:** 9691–9696.

Charles, M., et al. (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics **180:** 1071–1086.

Cossegal, M., Chambrier, P., Mbelo, S., Balzergue, S., Martin-Magniette, M.L., Moing, A., Deborde, C., Guyon, V., Perez, P., and Rogowsky, P. (2008). Transcriptional and metabolic adjustments in ADP-glucose pyrophosphorylase-deficient bt2 maize kernels. Plant Physiol. **146:** 1553–1570.

Devos, K.M., Brown, J.K., and Bennetzen, J.L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis.* Genome Res. **12:** 1075–1079.

Devos, K.M., Ma, J., Pontaroli, A.C., Pratt, L.H., and Bennetzen, J.L. (2005). Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc. Natl. Acad. Sci. USA **102:** 19243–19248.

Dolezel, J., Simkova, H., Kubalakova, M., Safar, J., Suchankova, P., Cihalikova, J., Bartos, J., and Valarik, M. (2009). Chromosome genomics in the Triticeae. In Plant Genetics and Genomics, C. Feuillet and G.J. Muehlbauer, eds (New York: Springer), pp. 285–316.

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

Erayman, M., Sandhu, D., Sidhu, D., Dilbirligi, M., Baenziger, P.S., and Gill, K.S. (2004). Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res. **32:** 3546–3565.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Feuillet, C., and Salse, J. (2009). Comparative genomics in the Triticeae. In Plant Genetics and Genomics, C. Feuillet and G.J. Muehlbauer, eds (New York: Springer), pp. 451–477.

Gill, B.S., Friebe, B., and Endo, T.R. (1991). Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum aestivum*). Genome **34:** 830–839.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: A graphical tool for sequence finishing. Genome Res. **8:** 195–202.

Griffiths, S., Sharp, R., Foote, T.N., 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.

Haberer, G., et al. (2005). Structure and architecture of the maize genome. Plant Physiol. **139:** 1612–1624.

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. USA **99:** 8133–8138.

Hufton, A.L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: A variety of outcomes. Curr. Opin. Genet. Dev. **19:** 600–606.

Huo, N., Vogel, J.P., Lazo, G.R., You, F.M., Ma, Y., McMahon, S., Dvorak, J., Anderson, O.D., Luo, M.C., and Gu, Y.Q. (2009). Structural characterization of *Brachypodium* genome and its syntenic relationship with rice and wheat. Plant Mol. Biol. **70:** 47–61.

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

Jaillon, O., et al; French-Italian Public Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature **431:** 569–573.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA **94:** 1872–1877.

Krattinger, S., Wicker, T., and Keller, B. (2009). Map-based cloning of genes in Triticeae (wheat and barley). In Plant Genetics and Genomics, C. Feuillet and G.J. Muehlbauer, eds (New York: Springer), pp. 337–358.

Kronmiller, B.A., and Wise, R.P. (2009). Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the rf1-associated region of maize. Plant Physiol. **151:** 483–495.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics **9:** 517.

Larkin, M.A., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics **23:** 2947–2948.

Li, Q., Li, L., Dai, J., Li, J., and Yan, J. (2009). Identification and characterization of CACTA transposable elements capturing gene fragments in maize. Chin. Sci. Bull. **54:** 642–651.

Li, W., and Gill, B.S. (2002). The colinearity of the Sh2/A1 orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the *triticeae.* Genetics **160:** 1153–1162.

Li, W., Zhang, P., Fellers, J.P., Friebe, B., and Gill, B.S. (2004). Sequence composition, organization, and evolution of the core *Triticeae* genome. Plant J. **40:** 500–511.

Liu, R., Vitte, C., Ma, J., Mahama, A.A., Dhliwayo, T., Lee, M., and Bennetzen, J.L. (2007). A GeneTrek analysis of the maize genome. Proc. Natl. Acad. Sci. USA **104:** 11844–11849.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:** 955–964.

Luo, M.C., et al. (2009). Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc. Natl. Acad. Sci. USA **106:** 15780–15785.

Ma, J., and Bennetzen, J.L. (2004). Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA **101:** 12404–12410.

Ma, J., and Bennetzen, J.L. (2006). Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. USA **103:** 383–388.

McFadden, E., and Sears, E. (1946). The origin of *Triticum spelta* and its free-threshing hexaploid relatives. J. Hered. **37:** 81–89 107–116.

Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F., and Wing, R.A. (2004). Sequence composition and genome organization of maize. Proc. Natl. Acad. Sci. USA **101:** 14349–14354.

Metzker, M.L. (2009). Sequencing technologies - The next generation. Nat. Rev. Genet. **11:** 31–46.

Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y., and Shinozaki, K. (2009). TriFLDB: A database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol. **150:** 1135–1146.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. **37:** 997–1002.

Ogihara, Y., et al. (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res. **33:** 6235–6250.

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

Paux, E., et al. (2010). Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol. J. **8:** 196–210.

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.

Paux, E., et al. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. Science **322:** 101–104.

Paux, E., Tamasloukht, M., Ladouce, N., Sivadon, P., and Grima-Pettenati, J. (2004). Identification of genes preferentially expressed during wood formation in Eucalyptus. Plant Mol. Biol. **55:** 263–280.

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, L.L., et al. (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics **168:** 701–712.

Rabinowicz, P.D., Citek, R., Budiman, M.A., Nunberg, A., Bedell, J.A., Lakey, N., O'Shaughnessy, A.L., Nascimento, L.U., McCombie, W.R., and Martienssen, R.A. (2005). Differential methylation of genes and repeats in land plants. Genome Res. **15:** 1431–1440.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: Sequence visualization and annotation. Bioinformatics **16:** 944–945.

Sabot, F., Guyot, R., Wicker, T., Chantret, N., Laubin, B., Chalhoub, B., Leroy, P., Sourdille, P., and Bernard, M. (2005). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. Mol. Genet. Genomics **274:** 119–130.

Safar, J., et al. (2004). Dissecting large and complex genomes: Flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J. **39:** 960–968.

Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T.J., Messing, J., and Feuillet, C. (2009). Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc. Natl. Acad. Sci. USA **106:** 14908–14913.

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

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43–45.

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

See, D.R., Brooks, S., Nelson, J.C., Brown-Guedira, G., Friebe, B., and Gill, B.S. (2006). Gene evolution at the ends of wheat chromosomes. Proc. Natl. Acad. Sci. USA **103:** 4162–4167.

Smith, D.B., and Flavell, R.B. (1975). Characterisation of the wheat genome by renaturation kinetics. Chromosoma **50:** 223–242.

Soderlund, C., et al. (2009). Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. PLoS Genet. **5:** e1000740.

Sonnhammer, E.L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167:** GC1–GC10.

Sorrells, M.E., et al. (2003). Comparative DNA sequence analysis of wheat and rice genomes. Genome Res. **13:** 1818–1827.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S., and Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. **19:** 2221–2230.

Wessler, S.R., Bureau, T.E., and White, S.E. (1995). LTR-retrotransposons and MITEs: Important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. **5:** 814–821.

Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. (2003). CACTA transposons in *Triticeae*. A diverse family of high-copy repetitive elements. Plant Physiol. **132:** 52–63.

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. **17:** 1072–1081.

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T., Graner, A., Ware, D., and Stein, N. (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics **9:** 518.

Wicker, T., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8:** 973–982.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21:** 1859–1875.

Yu, J., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92.

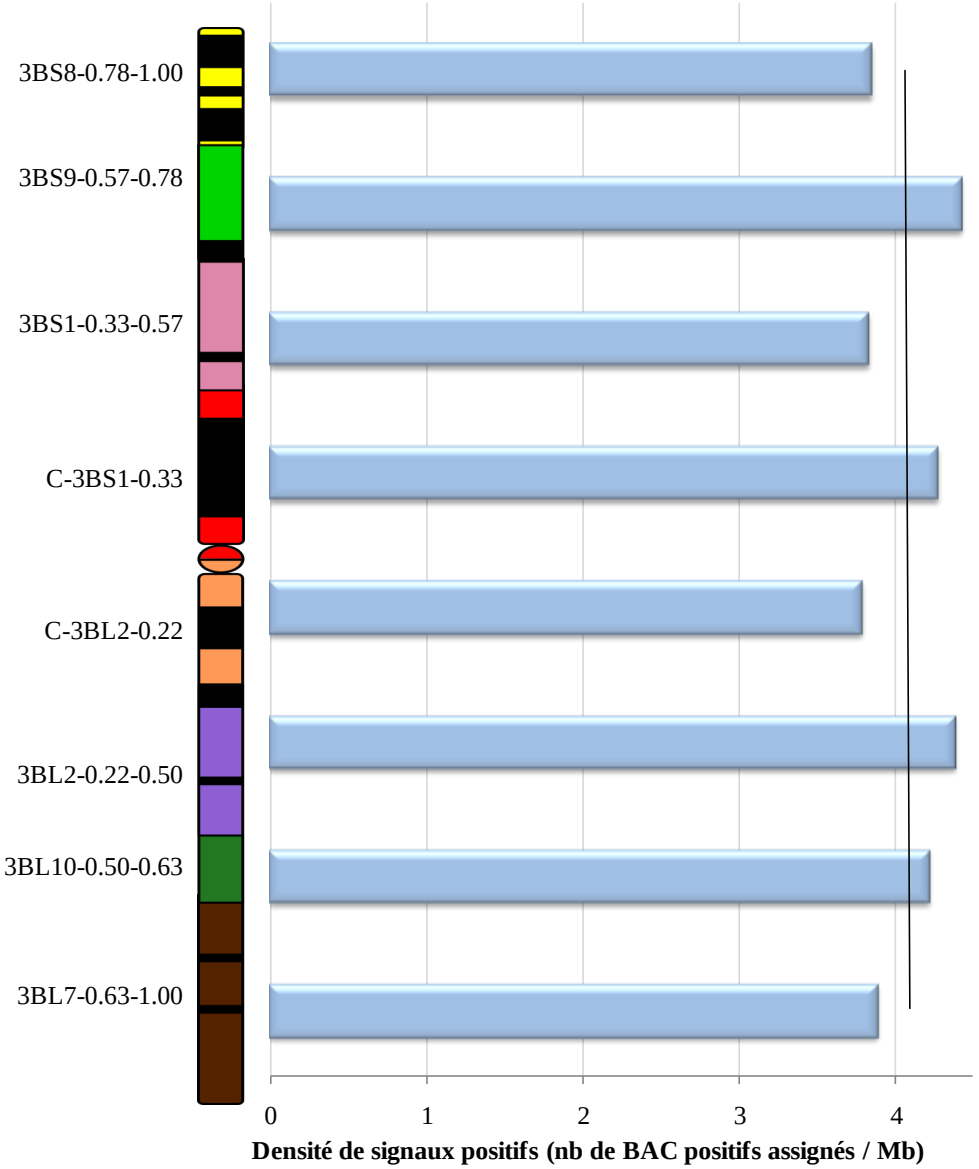


Figure 24 : Distribution uniforme de la densité de BAC portant au moins un gène le long du chromosome 3B.

Conclusion Article n°1

En hybridant les filtres portant les BAC du MTP du chromosome 3B avec des ADNc issus de cinq organes à différents stades de développement, nous avons pu obtenir une vision globale de l'organisation de l'espace génique à l'échelle du chromosome. Ainsi, nous avons pu montrer que l'espace génique couvrait l'ensemble du chromosome sans régions de plusieurs mégabases totalement dépourvues de gènes (Figure 24). Ces premiers résultats ont donc permis d'extrapoler à l'ensemble du chromosome les conclusions de l'étude de contigs séquencés sur l'organisation de l'espace génique chez le blé tendre. De plus, ils permettent d'infirmer l'hypothèse selon laquelle 94% des gènes seraient condensés dans les 29% les plus centromériques du génome de blé.

Cependant, si cette approche offre une vision globale du chromosome, elle ne permet pas d'analyser l'organisation de l'espace génique de façon précise notamment parce que le nombre de gènes par BAC ne peut pas être déterminé.

Pour approfondir ces résultats et aller plus avant dans notre étude, une nouvelle stratégie a été développée afin de parvenir à étudier les gènes d'une part à l'échelle du chromosome 3B entier mais également à l'échelle du BAC. Cette stratégie a consisté à hybrider les BAC du MTP du chromosome 3B poolés en trois dimensions sur une puce à ADN portant environ 15000 unigènes d'orge. Ce travail est le fruit d'une collaboration entre notre équipe et l'équipe de Robbie Waugh (Scottish Crop Research Institute, Invergowrie, Ecosse). Il a été soumis pour publication dans la revue BMC Genomics sous le titre « Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources ».

ARTICLE N°2

Mise en évidence de tendances spécifiques de l'organisation de l'espace génique chez le blé par la combinaison de ressources génomiques de blé et d'orge

Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources

Camille Rustenholz[1]*, Pete E Hedley[2]*, Jenny Morris[2], Frédéric Choulet[1], Catherine Feuillet[1], Robbie Waugh[2], Etienne Paux[1§]

[1]INRA UMR 1095, Génétique Diversité et Ecophysiologie des Céréales, 63100 Clermont-Ferrand, France

[2]Scottish Crop Research Institute, Invergowrie, Dundee, DD2 5DA, UK

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

CR: camille.rustenholz@clermont.inra.fr

PEH: pete.hedley@scri.ac.uk

FC: frederic.choulet@clermont.inra.fr

CF: catherine.feuillet@clermont.inra.fr

RW: robbie.waugh@scri.ac.uk

EP: etienne.paux@clermont.inra.fr

ABSTRACT

Background

Because of its size, allohexaploid nature and high repeat content, the wheat genome has always been perceived as too complex for efficient molecular studies. We recently constructed the first physical map of a wheat chromosome (3B). However gene mapping is still laborious in wheat because of high redundancy between the three homoeologous genomes. In contrast, in the closely related diploid species, barley, numerous gene-based markers have been developed. This study aims at combining the unique genomic resources developed in wheat and barley to decipher the organisation of gene space on wheat chromosome 3B.

Results

Three dimensional pools of the minimal tiling path of wheat chromosome 3B physical map were hybridised to a barley Agilent 15K expression microarray. This led to the fine mapping of 738 barley orthologous genes on wheat chromosome 3B. In addition, comparative analyses revealed that 68% of the genes identified were syntenic between the wheat chromosome 3B and barley chromosome 3H and 59% between wheat chromosome 3B and rice chromosome 1, together with some wheat-specific rearrangements. Finally, it indicated an increasing gradient of gene density from the centromere to the telomeres positively correlated with the number of genes clustered in islands on wheat chromosome 3B.

Conclusion

Our study shows that novel structural genomics resources now available in wheat and barley can be combined efficiently to overcome specific problems of genetic anchoring of physical contigs in wheat and to perform high-resolution comparative analyses with rice for deciphering the organisation of the wheat gene space.

BACKGROUND

The term "gene space" refers to the fraction of the genome corresponding to protein coding genes and, by extension, to the distribution of these genes (Jackson et al., 2004). In large genomes that contain abundant repetitive DNA, it encompasses also the notion of regions containing genes, the so-called gene-rich regions, surrounded by gene-poor regions composed of repeats (Varshney et al., 2006).

With the growing number of sequenced plant genomes, it becomes obvious that the distribution pattern of genes is far from random and not universal across the plant kingdom. Small plant genomes, such as *Arabidopsis thaliana* (125 Mb), *Brachypodium distachyon* (272 Mb) and *Oryza sativa* (389 Mb) exhibit fairly homogenous gene distribution along their chromosomes (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; The International *Brachypodium* Initiative, 2010). The transition from a homogenous to a non-homogenous gene distribution seems correlated to the genome size. Indeed, in intermediate size genome, such as *Populus trichocarpa* (485 Mb) and *Vitis vinifera* (487 Mb), large regions alternating between high and low gene density were observed (Tuskan et al., 2006; Jaillon et al., 2007), whereas larges genomes, such as *Glycine max* (1115 Mb) and *Zea mays* (2300 Mb), display an increasing gradient of gene density from the centromere to the telomeres (Schnable et al., 2009; Schmutz et al., 2010).

Because of its size (17000 Mb), allohexaploid nature (A, B and D-genomes) and high repeat content (>80%) (Zonneveld et al., 2005), the bread wheat (*Triticum aestivum* L.) genome is among the largest and most complex plant genomes and has always been considered too complex for molecular analyses. As a result, no genome sequence is available yet and very little is known about the organisation of the wheat gene space. The first insights were obtained from the mapping of wheat gene-based markers in wheat aneuploid genotypes called deletion lines where fragments of chromosomes or deletion bins are missing (Endo and Gill, 1996). Based on EST and *Pst1* genomic clone mapping, Erayman et al. (2004) suggested a very heterogeneous distribution of the genes along the wheat chromosomes, with 94% of the genes being located in only 29% of the entire wheat chromosomes and mostly at their telomeric parts. In contrast, by EST mapping on chromosome group 3 deletion bins, Munkvold et al. (2004) observed a slight gradient of the gene density along the chromosomes as well as a significant number of genes in the most proximal bins thereby suggesting a more

homogeneous distribution. More recently, individual BAC sequencing (Devos et al., 2005; Charles et al., 2008) confirmed a rather homogeneous gene distribution in wheat with an average of one gene per BAC. Finally, Choulet et al. (2010) investigated megabase-sized regions from various parts of chromosome 3B and indicated that the gene-free regions are much smaller than expected by Erayman et al.(2004), *i.e.* not larger than 1 Mb. Moreover, they found evidence for a slight gradient (twofold) of the gene density distribution from the centromere to the telomeres. Thus, additional whole genome or whole chromosome analyses are needed to better characterize the gene space organisation in wheat.

We recently constructed a physical map of chromosome 3B, the largest wheat chromosome (1 Gb, 2.5 times the whole rice genome) (Paux et al., 2008). The map consists of 1036 contigs spanning 811 Mb, of which 611 Mb are anchored with 1443 molecular markers. However, very few contigs are anchored by gene-derived markers. Indeed despite the development of genomic resources, such as extensive marker collections and saturated genetic maps (GrainGenes 2.0, http://wheat.pw.usda.gov/GG2; Lehmensiek et al., 2009; Paux and Sourdille, 2009), genetic mapping of genes in wheat is still hampered by the lack of polymorphism and the presence of the three homoeologous copies of each gene. As a result, no high density transcript genetic map is available. In contrast, several gene maps have been constructed for barley (Stein et al., 2007b; Potokina et al., 2008; Close et al., 2009; Sato et al., 2009; Chen et al., 2010b) (*Hordeum vulgare* L.) that diverged from wheat ~10-12 MYA (Dvorak and Akhunov, 2005; Chalupska et al., 2008) and belongs to the same tribe (Triticeae). With a size of 4.9 Gb (Zonneveld et al., 2005) and a repeat content of over 80% (Wicker et al., 2008), the diploid barley genome (2n =14) is very similar to the wheat subgenomes and several mapping studies have demonstrated a high collinearity between barley and wheat (Moore et al., 1995; Dubcovsky et al., 1996; Devos and Gale, 2000; Bennetzen and Ramakrishna, 2002; Dvorak and Akhunov, 2005; Chalupska et al., 2008).

Here, we wanted to explore the possibility of using barley transcript genetic maps as a surrogate to anchor and order the wheat physical contigs. BAC pools representing the minimal tiling path (MTP) of wheat chromosome 3B were hybridised onto barley expression microarrays to identify the location of genes along the wheat 3B physical map. The results show that such barley-wheat cross-hybridisations represent high-throughput cost-efficient approaches for anchoring genes on wheat physical maps and for performing comparative genomics studies between wheat and other grass genomes. In addition, the possibility to

locate genes precisely within BAC contigs that were anchored by other markers onto the chromosome 3B enabled us to gain new insights into the distribution of genes along a wheat chromosome.

RESULTS AND DISCUSSION

A high throughput anchoring method

To assess the efficiency of wheat-barley cross-species hybridisation for gene-based physical map anchoring, a barley Agilent 15K unigene microarray was hybridised with 60 three-dimensional (plate, row, column) BAC pools from the minimal tiling path (MTP) of the wheat chromosome 3B (Paux et al., 2008). After signal quantification and normalisation, hybridisation data were evaluated with four complementary scoring methods to reliably locate as many barley gene homologs as possible on the wheat BACs (see Materials and methods). Using the most stringent "automated scoring" method, 3355, 3401 and 3286 probes were identified as positive with the plate, row and column pools, respectively. Deconvolution of the pool data led to the identification of 571 unambiguous BAC addresses for 566 unigenes, defining 561 unique genomic loci and 5 duplicated loci. The less stringent "boxplot scoring" method led to the identification of 6205, 5103 and 6761 positive probes for the plate, row and column pools, respectively. With this method, 770 probes having unambiguous BAC addresses were identified, including 481 that were already identified with the "automated" method. Out of the 289 newly identified probes, we selected 86 probes (100 loci) that correspond to the most robust data (*i.e.* located on two to three overlapping BACs). Finally the "semi-automated" and the "manual scoring" methods added additional BAC addresses for 13 and 78 probes respectively that showed missing coordinates with the two other methods due to technical limitations (detailed below).

In total, the combination of four methods enabled us to identify 762 unambiguous wheat BAC addresses for 743 barley probes. A BLASTN search (Altschul et al., 1997) against the Triticeae repeat database TREP (Wicker et al., 2002), indicated that five probes had high sequence identity (> 86%) with TEs and were removed from further analysis. Each of the remaining 738 non TE-related genes was assigned to one to three wheat BACs resulting in 757 gene loci identified on the wheat chromosome 3B physical map (Paux et al., 2008). These

barley unigenes were located on 624 wheat BACs that corresponded to 388 individual contigs of 187 kb to 3.8 Mb and 86 singletons.

We tested the reliability of the 757 genes using the sequenced contigs available on chromosome 3B (Choulet et al., 2010). We found that 74% (23/31) of the genes located on the sequenced contigs through hybridisation gave a hit on the sequenced contigs after a BLASTN analysis. Out of these genes, 91% (21/23) matched a gene on the sequenced contigs at their expected location. Out of the 15208 unigenes on the barley Agilent microarray, 199 gave a hit after a BLASTN analysis against the sequenced contigs but were not located on a BAC through hybridisation. Only 15% of them matched a gene whereas 85% of them matched repeats or regions for which no annotation was found. Therefore the assignation of unigenes on BACs using cross-species hybridisations is a reliable technique that allows to identify a subset of probes that are enriched in genes and to limit the number of false-positive loci.

The 738 probes correspond to roughly 40% of the barley unigenes that were expected to be present on the wheat chromosome 3B physical map. Indeed, chromosome 3H accounts for approximately 14.8% of the barley genome (Suchánková et al., 2006; Mayer et al., 2009). Assuming a comparable gene density for all barley chromosomes, 2250 probes out of the 15208 unigenes are expected to be located on chromosome 3H. As the MTP covers 82% of the whole wheat chromosome 3B, about 1845 probes should in theory be present on the wheat chromosome 3B physical map assuming that all barley genes are conserved in wheat. The difference of 60% between expected and observed results could be explained by both biological and technical limitations of our experiment. First, sequence divergence between wheat and barley genes may have significantly impacted the efficiency of this approach. Letowski et al. (2004) estimated that hybridising a probe and a DNA target sharing 90% of sequence identity results in 73% to 99% decrease in hybridisation signal intensity compared to a probe and a DNA sharing 100% of sequence identity. Blasting the barley 60-mer probes against the 6162 wheat *cv.* Chinese Spring full-length cDNA dataset (Kawaura et al., 2009) revealed that 56% of the hits show more than 10% nucleotide divergence (86% identity on average). Moreover we found that the unigenes located on the sequenced contigs of the wheat chromosome 3B (Choulet et al., 2010) through BLASTN and hybridisation showed a significantly higher sequence identity (90%) compared to the ones that were located on the sequenced contigs through BLASTN only (83%) (T-test, *P*-value=5E-6). Therefore, one can estimate that sequence conservation played a key role in the detection of hybridisation signals

and that more than half of the potentially positive barley probes generated a near undetectable hybridisation signal with the wheat BACs. A second origin of the discrepancy likely originates from the presence of gene families located at multiple loci. The wheat genome is allohexaploid (three subgenomes: A, B and D) and at least one copy of each wheat gene is expected to be present on the three homoeologous chromosomes. In addition, there is increasing evidence for high level of tandem and interchromosomal duplication events in wheat and perhaps barley genomes since their divergence from the other grasses (Akhunov et al., 2003; Choulet et al., 2010). Thus there is a good probability that some genes are found in multiple copies on chromosome 3B. Such genes can result in multiple non-overlapping BAC addresses that cannot be resolved without ambiguity and are therefore excluded from our analysis. Another critical point affecting the efficiency of the approach lies in the putative heterogeneity of the BAC pools. Indeed, each three-dimensional MTP pool contains more than 300 BACs (see Methods), making it difficult to guarantee equimolarity for all BACs. In some extreme cases, this heterogeneity in individual BAC quantity may lead to weak signal intensity for positive probes resulting in missing coordinates. These two limitations could be circumvented by the use of six-dimension pools of the complete chromosome 3B BAC library (Klein et al., 2000). However, such pools would have required almost 3 times more hybridisations than the three-dimensional MTP pools (177 vs. 60) thereby reducing the cost-efficiency of the approach.

Despite these limitations, this single experiment permitted the localisation of 738 genes on the wheat chromosome 3B contigs and allowed us to get novel information for the order of the BAC contigs along the chromosome based on the barley EST genetic maps. So far, genetic mapping of genes in wheat has been hampered by the lack of polymorphism in the genic sequences and the presence of several homoeologous copies. As a consequence, only a third of the 680 markers located on the chromosome 3B genetic map constructed using the ‘neighbours’ approach correspond to ESTs (Paux et al., 2008). Here, we established a barley whole genome neighbour map using the same criteria as the IBM neighbour map of maize (Cone et al., 2002) and used it to assess the order of wheat contigs based on the EST order found on the genetic map. Out of the 738 probes assigned to BAC contigs of wheat chromosome 3B, 308 (42%) were mapped to the barley neighbour map including 209 on chromosome 3H and 99 on other chromosomes. Using the barley 3H mapping data, 151 BAC contigs and 20 singletons from the wheat chromosome 3B physical map were genetically

ordered. Only 30% of these contigs were previously ordered genetically using the wheat chromosome 3B neighbour map, whereas 44% were mapped to the wheat chromosome 3B deletion bin map, but not ordered in bins, and 26% were not anchored at all (Paux et al., 2008). In addition, it is worth noting that 36% of the 151 contigs are only anchored by gene-based markers. This is consistent with the results of Paux et al. (2008) who showed that some regions of the genome can only be anchored by specific types of markers (ESTs, SSRs, ISBPs) and that 35% of the contigs were anchored by ESTs only. Therefore, we conclude that the cross-hybridisations of wheat BAC pools with barley expression microarrays is a straight forward approach to order wheat contigs with gene-based markers without the difficulty of EST genetic mapping in wheat.

Moreover, the total cost for these 60 pool hybridisations on 15 microarrays was approximately 8800 USD. For the same price, PCR screening of individual EST markers on the same BAC pools (including primers and amplification) would only have allowed testing of 500 markers. Thus, the method is a cost-efficient alternative to PCR-based physical map anchoring. However, despite its convenience and its cost-efficiency, this technique is still limited in the number of contigs anchored and ordered but it would be greatly improved by technological developments in a near future. First, use of the barley 44K Agilent expression microarray will significantly increase the number of positive probes, regardless of the experiment efficiency. Second, as large amounts of barley SNPs are becoming available (Close et al., 2009), the number of genetically mapped genes will increase in the coming years, therefore improving the efficiency of the anchoring strategy.

Finally, wheat-barley cross-species hybridisation is a convenient, cost-efficient and relatively high-throughput approach for gene-based physical map anchoring and ordering of wheat BAC contigs. However, even if the use of barley genomic resources circumvents the limitations caused by the complexity of the wheat genome, the divergence between the two species is large enough to observe synteny breaks. Thus, we performed a comparative study between wheat, barley and rice to assess the extent to which the barley gene order is transferable to wheat.

Table 3 : Mapping data in wheat chromosome 3B deletion bins.

Wheat				**Barley**			**Rice**		
3B Deletion Bin	Bin size (Mb)	Number of loci	Density (locus / Mb)	Genes mapped on 3H	Genes mapped on the other chromosomes	Collinear genes between 3B and 3H	Genes mapped on Os01	Genes mapped on the other chromosomes	Collinear genes between 3B and Os01
3BS8-0.78-1.00	44.2	37	0.84	10	9	7	15	14	10
3BS9-0.57-0.78	43.2	43	1.00	12	4	9	22	16	16
3BS1-0.33-0.57	94.3	86	0.91	14	11	8	40	33	20
C-3BS1-0.33	58.3	34	0.58	9	1	4	23	7	16
C-3BL2-0.22	45.7	36	0.79	15	3	8	27	6	18
3BL2-0.22-0.50	74.9	78	1.04	25	10	21	47	26	38
3BL10-0.50-0.63	40.1	40	1.00	13	5	8	18	17	8
3BL7-0.63-1.00	155.5	165	1.06	55	14	37	93	57	59
Total assigned	**556.2**	**519**	**0.93**	**153**	**57**	**102**	**285**	**176**	**185**
Not assigned	438.8	238	/	66	42	/	111	104	/
Total	995	757	/	219	99	/	396	280	/
				318			676		

Comparative genomics between wheat, barley and rice

In addition to its interest for anchoring physical maps, cross-species hybridisation also provides valuable data for comparative genomics as it allows the mapping of barley (and to some extent rice) orthologous genes on wheat chromosomes. We studied synteny, *i.e.* the conservation of the genes on the orthologous chromosomes of wheat chromosome 3B in barley (chromosome 3H) and rice (chromosome 1) without the assumption of the conservation of the gene order (Keller and Feuillet, 2000). Out of the 738 probes located on the contigs of wheat chromosome 3B, 209 were mapped on barley chromosome 3H and 99 on other barley chromosomes. Rice orthologous genes were identified unambiguously for 659 of the 738 probes located on wheat chromosome 3B, of which 389 are located on rice chromosome 1and 270 on other rice chromosomes (Table 3). These results suggest that, at the whole chromosome scale, 68% and 59% of wheat chromosome 3B genes are syntenic with genes located on the orthologous barley chromosome 3H and rice chromosome 1, respectively. These results are consistent with previous studies that estimated between 59% and 74% of the genes were in conserved positions between wheat chromosome group 3 and rice chromosome 1 and 75% between wheat chromosome group 3 and barley chromosome 3H (Gaut, 2002; La Rota and Sorrells, 2004; Munkvold et al., 2004; Devos, 2005; Varshney et al., 2005; Cho et al., 2006; Bilgic et al., 2007; Stein et al., 2007b; Bolot et al., 2009; Thiel et al., 2009). The genes that are not syntenic between wheat chromosome 3B and barley chromosome 3H mapped on the other barley chromosomes with no significant bias towards any other single chromosome (Chi² test, *P*-value=0.16). In contrast, the non-syntenic genes between wheat chromosome 3B and rice chromosome 1 are biased in favour of genes mapped on the rice chromosomes carrying the highest number of genes (chromosomes 3 and 5) and against the rice chromosome carrying the lowest number of genes (chromosome 9) (Chi² test, P-value$<10^{-5}$). Further analyses confirmed that the distribution of the non-syntenic genes between wheat chromosome 3B and rice chromosome 1 on the other rice chromosomes is correlated with the number of genes per rice chromosome (Pearson's correlation coefficient r=0.735; *P*-value=0.01) (International Rice Genome Sequencing Project, 2005). Thus, no real mapping bias was identified towards any of the non-syntenic barley or rice chromosomes.

Interestingly, a number of genes located on wheat chromosome 3B were not syntenic with barley chromosome 3H but their homologs were syntenic between barley and rice. For example, 11 wheat chromosome 3B genes mapped on barley chromosome 2H and on its

ortholog in rice (chromosome 4).We found another example with 9 wheat chromosome 3B genes mapping on barley chromosome 6H and on the orthologous rice chromosome 2 (Bolot et al., 2009). This result indicates that these genes have undergone rearrangements specifically in wheat and supports the recent finding of Choulet et al. (2010) for extensive interchromosomal duplications in wheat.

Out of the 209 probes mapped on barley chromosome 3H, 153 have been assigned to one of the eight deletion bins of wheat chromosome 3B and therefore, their approximate location on the chromosome arms is known. This enabled us to study the synteny between wheat, barley and rice at a finer scale. We calculated the percentage of probes that are syntenic to barley chromosome 3H genes for each deletion bin of chromosome 3B and found that the conservation of genes is significantly uniform along chromosome 3B (Chi^2 test, *P*-value=0.84) with 73% of syntenic genes per bin on average (Table 3). We performed the same calculation with the 285 genes assigned to wheat chromosome 3B deletion bins and syntenic to genes on rice chromosome 1. In this case, the distribution of syntenic genes was negatively correlated with the distance to the centromere (Pearson's correlation coefficient r=-0.742; *P*-value=0.04). In other words, the level of synteny between wheat chromosome 3B and rice chromosome 1 decreases from the centromere to the telomeres. This is in complete agreement with the results of Akhunov et al. (2003) who correlated this with the recombination rate along wheat chromosomes. However, using the data from Saintenac et al. (2009) who performed an analysis of the distribution of the recombination rate among chromosome 3B, we did not find any correlation between the synteny level and crossing-over frequency (Pearson's correlation coefficient r=-0.378; *P*-value=0.36). Comparisons between the sequences of 13 Mb sized contigs of chromosome 3B with the rice and *Brachypodium* genomes led to the same conclusions (Choulet et al., 2010). Moreover, the authors found a positive correlation between transposable element activity and the number of non syntenic genes. Thus, it is likely that the synteny level between wheat chromosome 3B and rice chromosome 1 that decreases from the centromere to the telomeres results from a combination of factors that have still to be identified.

The links between the barley chromosome 3H genetic map, the rice chromosome 1 sequence and the wheat chromosome 3B deletion bin map were used to analyse the collinearity, *i.e.* the order of the genes (Keller and Feuillet, 2000), between the genes on wheat chromosome 3B and on barley chromosome 3H and between the genes on wheat chromosome 3B and on rice

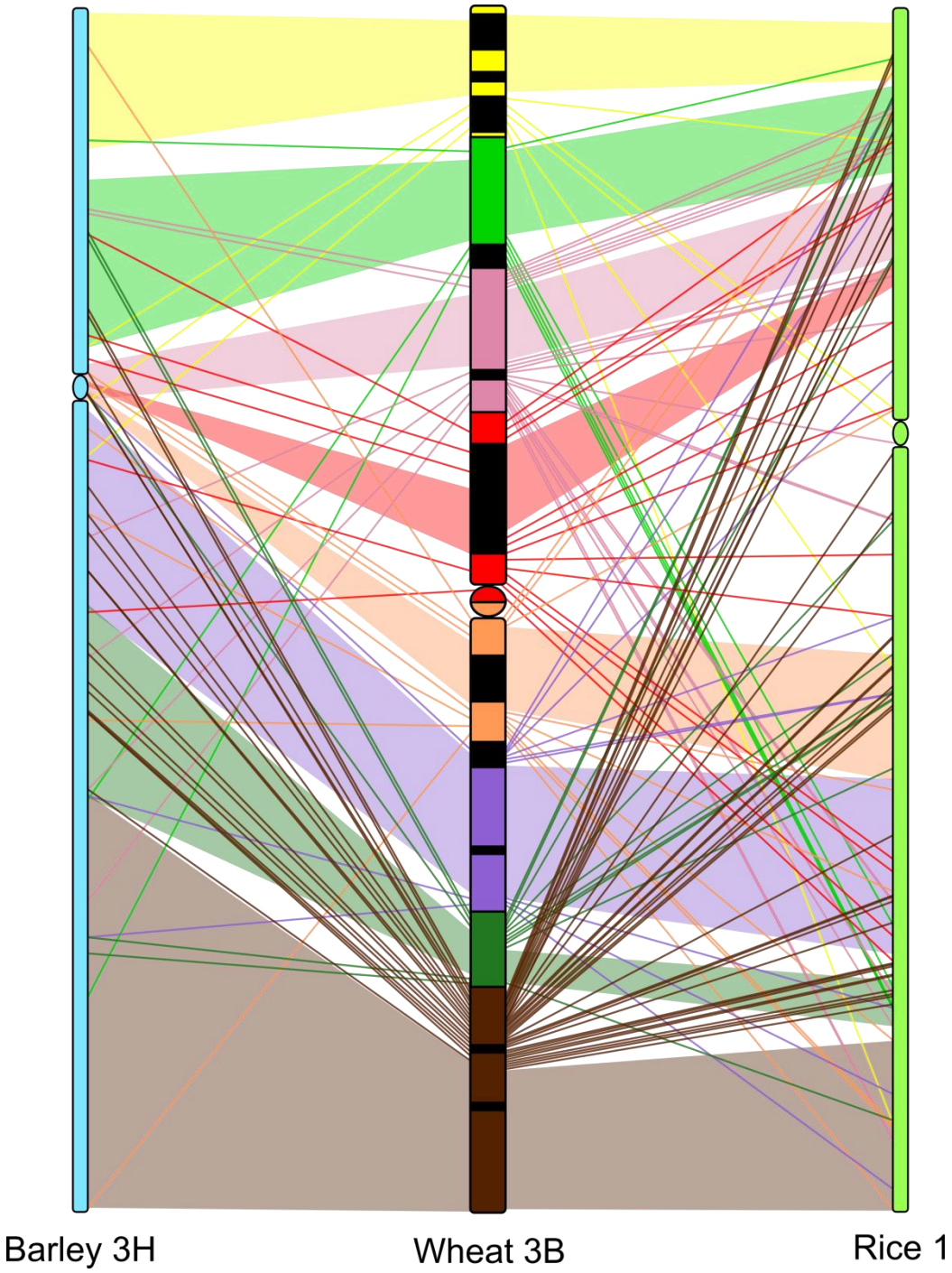


Figure 25 : Collinearity between wheat chromosome 3B, barley chromosome 3H and rice chromosome 1. Each colour on wheat chromosome 3B corresponds to a deletion bin. Yellow: 3BS8-0.78-1.00 ; light green: 3BS9-0.57-0.78 ; pink: 3BS1-0.33-0.57 ; red: C-3BS1-0.33 ; orange: C-3BL2-0.22 ; purple: 3BL2-0.22-0.50 ; dark green: 3BL10-0.50-0.63 and brown: 3BL7-0.63-1.00. The black segments correspond to the heterochromatic regions identified by C-banding, the coloured segments to the euchromatic regions and the circle to the centromere. The coloured blocks represent regions where the genes are collinear between two chromosomes. The lines represent the genes that are not collinear between two chromosomes. As the relative order of wheat genes in a given deletion bin is not known, we inferred this order from the barley chromosome 3H and from rice chromosome 1 data.

chromosome 1. As the relative order of wheat genes in a given deletion bin is not known, we inferred this order from the barley chromosome 3H and rice chromosome 1 data. In total we found that 102 (67%) genes were collinear between wheat 3B and barley 3H and 185 (65%) between wheat 3B and rice 1 (Table 3 & Figure 25). We calculated the percentage of collinear genes between wheat chromosome 3B and barley chromosome 3H and between wheat chromosome 3B and rice chromosome 1 for each bin. We found that the distribution of collinear genes is significantly uniform along chromosome 3B (Chi² test, *P*-value=0.89 and 0.69 for barley and rice respectively). Some translocations of genes can be observed between wheat chromosome 3B and barley chromosome 3H and between wheat chromosome 3B and rice chromosome 1 (Figure 25) that are similar to previous studies (Munkvold et al., 2004; Salse et al., 2008a). However, we expected a higher collinearity between the wheat and barley group 3 chromosomes with regard to previous results (Moore et al., 1995; Dubcovsky et al., 1996; Devos and Gale, 2000; Bennetzen and Ramakrishna, 2002). We think that this apparent discrepancy may originate from the construction of the barley neighbour map. None of the five individual barley genetic maps available to date holds a sufficient number of genes, and therefore we had to use a barley neighbour map combining these maps to optimize the anchoring experiment. However, the gene order is not fully reliable in such maps especially in the peri-centromeric and centromeric parts of the chromosomes where recombination is reduced or totally suppressed (Stein et al., 2007b). Finally, the use of wheat neighbour genetic mapping data suggests additional rearrangements in bins (data not shown) as suggested by Liu et al. (2005) and Chantret et al. (2008) and this may also lead to some discrepancies.

Altogether, our results regarding conservation between wheat chromosome 3B, barley chromosome 3H and rice chromosome 1 at the whole chromosome and at the deletion bins scales are in agreement with previous studies. However, we also noticed some wheat-specific rearrangements of the genes that disrupt the collinearity between wheat and barley and between wheat and rice. Thus, globally we expected the genes to be in the same order between wheat and barley but rearrangements are likely to be observed locally. So the results of anchoring and ordering of the wheat BAC contigs along chromosome 3B using the barley mapping data should be considered with caution as they may not be perfectly exact.

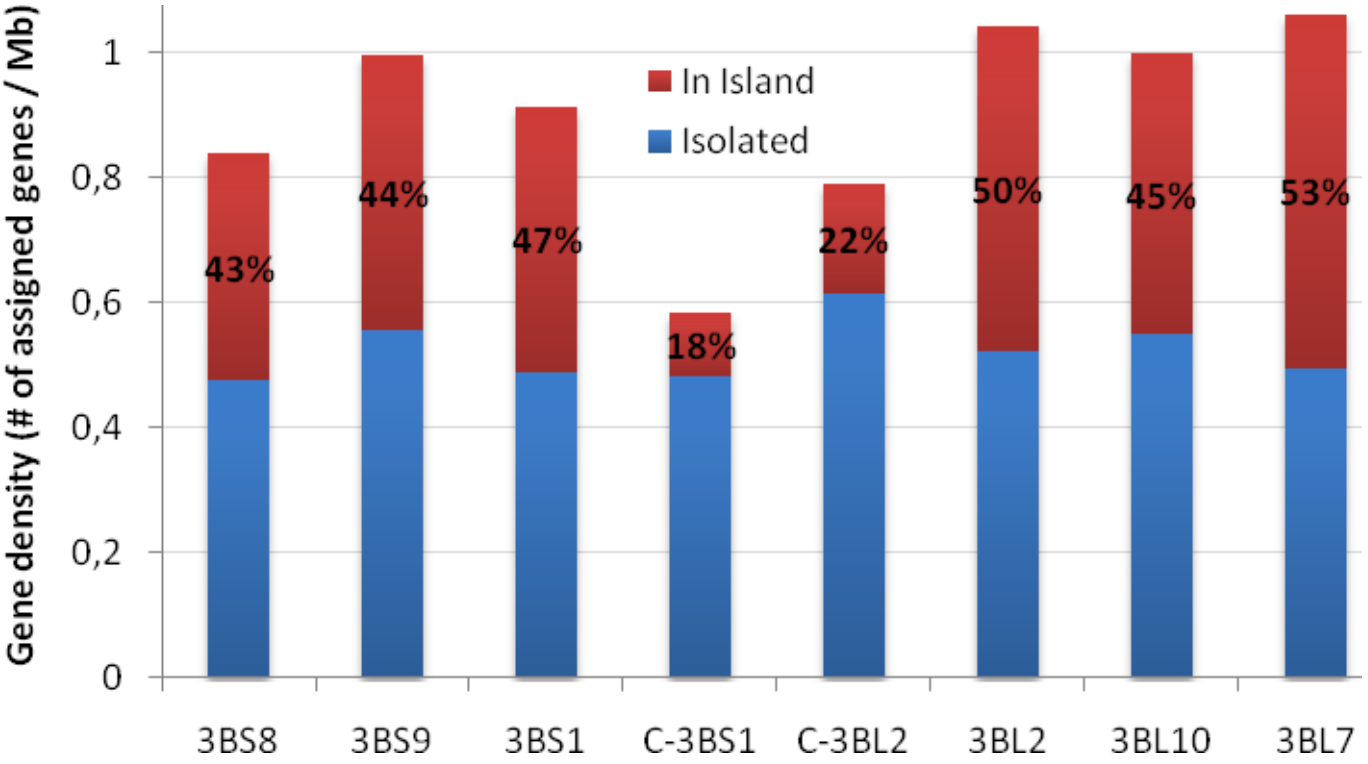


Figure 26 : Gene density in the eight deletion bins of wheat chromosome 3B. The density of isolated genes is represented by blue bars. The density of the genes organized in island is represented by red bars. The proportions of genes organized in island per deletion bins are shown as percentages within the red bars.

Wheat gene space organisation

For the first time, we were able to assign precisely a large number of genes to individual BACs and BAC contigs whose order is known on wheat chromosome 3B. This led us to analyse the pattern of gene distribution along the chromosome. Out of the 757 loci mapped on the wheat chromosome 3B physical map, 519 loci were assigned to the eight deletion bins (Table 3). The density of loci per deletion bin was calculated by dividing the number of loci assigned to each deletion bin by the cumulative length of contigs in the bin (Paux et al., 2008). The density of loci showed a slight increasing gradient from the centromere to the telomeres (Figure 26) with a positive correlation with the distance to the centromere (Pearson's correlation coefficient r=0.664), even though this correlation was not statistically significant using a 5% threshold (*P*-value=0.07). The highest density was found on the most distal 3BL7-0.63-1.00 deletion bin of the long arm (1.06 probes per megabase), whereas the lowest density was observed on the most proximal C-3BS1-0.33 bin of the short arm (0.58 probes per megabase). Such a gradient could be artificially created by a difference in gene sequence conservation from the centromere to the telomeres between wheat and barley. To test this hypothesis, we calculated the coefficient of correlation between the gene density and the percentage of sequence identity per bin. No significant correlation was found (Pearson's correlation coefficient r=-0.478; *P*-value=0.23), demonstrating that the gene density gradient is not biased by the differences in the similarity between wheat and barley gene sequences. Thus, the gene density distribution observed here provides a reliable snapshot of the gene space organisation along wheat chromosome 3B and led to a hypothesis where the gene density is higher in the distal parts than in the proximal parts of the chromosome (Figure 26). As we found that 87% of the genes were mapped on the 81% most distal parts of chromosome 3B, our result is in agreement with the moderate gradient of gene density along wheat chromosomes suggested by Munkvold et al. (2004), Devos et al. (2005), Charles et al. (2008) and more recently, Choulet et al.(2010). The discrepancy between these results and the first suggestions of Erayman et al. (2004) that most of the genes are located in the distal parts of the wheat chromosomes may originate from their use of a consensus deletion bin map from the A, B and D chromosomes, where markers were non-systematically assigned to the deletion bins (Saintenac et al., 2009). This likely led to approximations of the relative positions of the markers and the postulation that genes are mostly found in gene-rich regions in wheat. Here, the possibility to assign genes on physical contigs that cover 82% of a specific

chromosome 3B (Paux et al., 2008) enabled us to derive more precise information about the gene distribution.

We then extrapolated the expected gene density per deletion bin to the whole set of chromosome 3B genes. We first estimated the number of genes per bin by considering the bins fully covered by contigs and by keeping the same gene density distribution along chromosome 3B. This resulted in an estimate of 904 loci assigned to the eight deletion bins compared to the 519 loci identified by hybridisation in this study. Recently Choulet et al. (2010) estimated that chromosome 3B carries 8400 genes. We then extrapolated the gene density by considering 8400 loci assigned to the eight deletion bins. We found that the distal bin 3BL7-0.63-1.00 and the proximal bin C-3BS1-0.33 would have a gene density of 1 gene per 101 kb and 1 per 185 kb, respectively. Therefore, even in the least gene-dense regions of the chromosome, our results indicate that there may be one gene on average every 185 kb and therefore no megabase-sized regions devoid of genes. This is consistent with RNA hybridisations on chromosome 3B MTP arrays that showed that the largest region without genes is about 800 kb long and that genes are distributed across the entire chromosome 3B (Choulet et al., 2010).

However, our approach suffers from a major limitation to estimate the gene density precisely. Here, the gene densities estimated for the 3BS8-0.78-1.00 and 3BL7-0.63-1.00 telomeric deletion bins were lower than the ones found through RFLP hybridisation with ESTs by Munkvold et al. (2004) (normalized gene densities: 0.928 *versus* 1.190 and 1.176 *versus* 1.421, respectively). One of the characteristics of telomeric parts of wheat chromosomes is that they accumulate tandemly duplicated genes at a high rate (Akhunov et al., 2003; Choulet et al., 2010). Thus, it is likely that the differences in gene density observed between the two experiments reflect the inability of gene mapping based on BAC hybridisation to detect tandemly duplicated genes. This method is only qualitative and detects the presence or absence of a gene on a BAC but it does not indicate whether a gene located on a BAC is present in single or multiple copies. Thus, the gene density established through gene mapping based on BAC hybridisation is likely underestimated in the distal regions and therefore one can expect an even higher gene density gradient. If we consider that the difference in gene density between the two studies is only due to tandemly duplicated genes, we could estimate that we missed 28% and 21% of genes for 3BS8-0.78-1.00 and 3BL7-0.63-1.00 deletion bins, respectively. This also explains why our estimation of gene density at telomeres was lower

compared to the gene density of sequenced contigs located in distal regions of chromosome 3B (1 gene per 101 kb *versus* 1 gene per 86 kb) whereas our estimation at the centromere precisely fits the gene density of sequenced contigs located in proximal regions (1 gene per 185 kb *versus* 1 gene per 184 kb). Assuming that we missed 21% of genes due to tandem duplications in 3BL7-0.63-1.00 deletion bin, the gene density would be 1 gene per 90 kb. This demonstrates that all these studies can give an indication of the general gene space organisation along wheat chromosomes but are unable to precisely estimate the local gene density of specific regions.

To further study the gene space and especially the genes clustered in islands in more detail, we considered gene islands as multiple genes located on the same BAC or overlapping BACs, *i.e.* separated by less than 150 kb. Out of the 757 loci mapped on wheat chromosome 3B physical map, 303 loci, *i.e.* 40%, were considered part of gene islands, whereas the 454 remaining genes (60%) were considered as isolated genes. In contrast to the distribution of isolated genes that we found significantly uniform along chromosome 3B (Chi² test, *P*-value=0.97), the distribution of genes organised in islands was significantly non-uniform along chromosome 3B (Chi² test, *P*-value$<10^{-5}$) with a positive correlation between the density of genes in islands and the distance to the centromere (Pearson's correlation coefficient r=0.762; *P*-value=0.03) (Figure 26). We also found a correlation between the density of genes in islands and the overall gene density (Pearson's correlation coefficient r=0.956, *P*-value$<10^{-3}$). This strongly suggests that the gradient of gene density between centromeric and telomeric regions is due to the differential distribution of genes organised in islands across the chromosome with proportionately more genes in islands in the distal parts compared to the proximal parts.

In conclusion, our cross-species hybridisation technique allowed us to assign a large number of genes onto wheat chromosome 3B at the BAC resolution and to obtain original results on the wheat gene space organisation. We confirmed that the gene density distribution along the chromosome 3B follows a slight gradient from the centromere to the telomeres and we suggest that the presence of more gene islands in the distal part of the chromosome explains this gradient. However, the ultimate experiment to access the whole set of genes and confirm the gene density distribution at a high resolution along a wheat chromosome will be high-quality sequencing and annotation. This is currently underway for chromosome 3B (C. Feuillet, personal communication).

CONCLUSIONS

Our study demonstrates that hybridisations of the barley Agilent 15K expression microarray with wheat chromosome 3B MTP pools is a convenient and cost-efficient technique to perform physical map anchoring with gene-based markers. Our comparative genomics analysis between wheat, barley and rice confirms good global collinearity between these species, with a few wheat-specific rearrangements that could lead to local mis-ordering of wheat contigs using the barley gene order. Using this technique, we also confirmed previous studies that the gene space organisation follows a gradient of gene density along chromosome 3B from centromere to telomeres without large "gene-free" regions. We also demonstrated that this gradient was generated by a differential accumulation of gene islands between the centromere and the telomeres with more genes in islands in the distal parts of the chromosome. Such results have far-reaching implications in terms of strategies to sequence the wheat genome. Indeed, our results confirm that to access the whole wheat gene set, the entire wheat genome needs to be sequenced. A wheat expression microarray is currently being utilised to increase the density of genes at the BAC scale located along wheat chromosome 3B and to improve our understanding of the wheat gene space organisation.

METHODS

Barley expression microarray and hybridisations

The barley Agilent 15K expression microarray contains 15208 barley 60-mer probes derived from unigenes of HarvEST assembly #25 used to originally design probe sets for the 22K Barley1 Affymetrix GeneChip (Chen et al., 2010b). BACs (7440 in total) arranged in twenty 384-well plates were selected to build a wheat chromosome 3B Minimal Tiling Path covering 82% of the whole chromosome with ~30% overlap as described by Paux et al. (2008). These twenty plates were pooled in three dimensions (20 plate pools, 16 row pools and 24 column pools) to generate 60 samples by CNRGV (Toulouse, France) and the BAC pools were amplified as described by Paux et al. (2008). Two channels processing of the microarrays was used, with BAC pool DNA labelled with Cy3 and a mixed reference set of barley cv. Golden Promise RNAs (equal amounts of leaf, root and inflorescence) labelled with Cy5. RNA (5 µg) was labelled as described by Ducreux et al. (2008). Amplified BAC pool

DNA (200 ng) was labelled using a modified BioPrime Genomic DNA Labelling System (Invitrogen, Carlsbad, California USA): BAC pool DNA in 11 µl was added to 10 µl Random Primer Reaction Buffer mix and denatured at 95°C for 5 min prior to cooling on ice and to this was added 2.5 µl modified 10x dNTPs buffer (1.2 mM each of dATP, dGTP, dTTP; 0.6 mM dCTP; 10 mM Tris pH8.0; 1 mM EDTA), Cy3 dCTP (1 µl of 1 nM) and 0.5 µl Klenow enzyme (20U) followed by incubation for 16 h at 37°C. Labelled samples (BAC DNA & reference RNA) for each array were combined and unincorporated dyes removed using the Qiaquick PCR Purification Kit (Qiagen, Hilden, Germany) as recommended, eluting with 20 µl EB buffer (Qiagen, Hilden, Germany). Hybridisations and washing were carried out as recommended (Agilent Protocol v5.5). Scanning was performed with an Agilent G2505B scanner using default settings and data extracted using Agilent FE software (v 9.5.3). All data has been submitted to ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) (accession # E-TABM-1011) under MIAME guidelines (http://www.mged.org/Workgroups/MIAME/miame.html).

Blast analyses

A BLASTN analysis (Altschul et al., 1997) was performed with the 60-mer barley probes against the TREP database (Wicker et al., 2002) to identify probes that could hybridise with TEs of the wheat BACs. We considered that a probe could generate a false positive due to TEs if we found 80% identity on a minimum 45 nucleotides. Then a BLASTN analysis (Altschul et al., 1997) was performed with the 15208 60-mer barley probes against the sequenced contigs of wheat chromosome 3B (Choulet et al., 2010). The annotation of the best hit on the sequenced contigs was viewed using Artemis (Rutherford et al., 2000). The best hit and the query barley probe were then aligned using ClustalW2 (Larkin et al., 2007) and the sequence identity was calculated using the entire barley probe length. In addition, a BLASTN analysis (Altschul et al., 1997) was performed with the 60-mer barley probes against 6162 wheat *cv* Chinese Spring full-length cDNAs developed by Kawaura et al. (2009). The sequence identity between the best hit and the query barley probe was calculated on the entire barley probe length as previously described. The most significant rice homologues to the unigenes used to design the barley microarray probes were identified by BLASTN searches of

the gene models from the Rice Genome Annotation Project from Michigan State University (http://rice.plantbiology.msu.edu/, Rice Pseudomolecules v5 database).

Data deconvolution

Following hybridisation, signals were analysed to rebuild the MTP addresses of the BACs carrying an ortholog of a barley probe. For each barley probe, we identified the positive pools to determine the original MTP BAC address on which it is located. Each type of pool does not contain the same number of BACs (plate: 384 BACs/pool; row: 480 BACs/pool; column: 320 BACs/pool).

The first normalisation step undertaken addressed the fact that the 60 samples had different hybridisation signal averages. Medians were calculated for each pool independently and each value was divided by the median corresponding to the pool type. This led to comparable hybridisation values for each pool. A second normalisation step was undertaken for each probe, based on the same method. After this second normalisation step, probe hybridisation values were all comparable, while pool hybridisation values were not significantly changed.

To identify pools with positive signal, we first used an automated classical outlier detection method, that we called the "automated scoring" method. The mean and the standard deviation were calculated for each probe and used to define a different threshold for each probe. Calculation of the thresholds was different for each pool type (plate: Mean + 2.8 × Standard Deviation; row: Mean + 2.5 × Standard Deviation; column: Mean + 3 × Standard Deviation). All the pools with probe signal above this threshold were considered positive. We repeated this step twice by deleting positive signals previously detected, calculating the mean, standard deviation and the threshold again for each probe and selecting the new positive signals above the new thresholds. The calculation of the thresholds for the three pool types remained the same.

Following this "automated scoring" method, a "semi-automated" one was performed to identify missing coordinates from probes having five positive pools (*e.g.* two plate coordinates, two row coordinates and one column coordinate). Here, a combination of all possible coordinates was used to try to identify two overlapping BACs. A "manual scoring"

analysis was also performed to identify the missing coordinate from probes having two positive pools.

The final analysis that we called "boxplot scoring" method was performed whereby a boxplot was drawn using R software (www.r-project.org) for each probe for each of the three pool types and the upper outlier values were considered as positive pools. This analysis was less stringent than the "automated scoring" so we kept only the probes that were located on two overlapping BACs.

To rebuild the BAC addresses, we collated the positive pools for each probe. Some probes gave one positive pool per pool type which enables us to identify an unambiguous BAC address. However, some probes gave more positive pools per pool type. We therefore looked at every combination and used the chromosome 3B physical map data (Paux et al., 2008) to assist finding addresses of overlapping BACs where the probe is located on the overlap itself.

After identification of BACs carrying barley orthologs, we used the physical map (Paux et al., 2008) to locate them in their respective contig and possibly in one of the eight chromosome 3B deletion bins used for this study.

Synteny and collinearity analyses

In total, five barley genetic maps (Stein et al., 2007b; Potokina et al., 2008; Close et al., 2009; Sato et al., 2009; Chen et al., 2010b) were used to establish a barley neighbour map using the same criteria as the IBM neighbour map of maize (Cone et al., 2002). For two maps (Stein et al., 2007b; Sato et al., 2009), we had to perform a BLASTN analysis to link the markers to the barley unigenes mapped on wheat chromosome 3B. The best hits with at least 85% identity over 100 nucleotides were selected for each unigene. The order of the rice genes along the chromosomes was established using the rice gene numbering annotation. As the relative order of wheat genes in a given deletion bin is not known, we inferred this order from the barley chromosome 3H and rice chromosome 1 data. The software GenomePixelizer (http://www.atgc.org/GenomePixelizer/GenomePixelizer_Welcome.html) was used for the graphical display of the collinearity between wheat chromosome 3B, barley chromosome 3H and rice chromosome 1.

Statistical analyses

The statistical analyses including the T-test, Chi² and Pearson's correlation coefficient tests were performed using R software (www.r-project.org) at a 5% threshold. The distance to centromere was estimated from the centromere to the middle of deletion bins. We calculated the gene density per deletion bins by dividing the number of genes assigned to the bin by the length of the contigs assigned to the same bin. The gene density per bin of Munkvold et al. (2004) was estimated by dividing the number of genes they assigned to the bin by the total size of the bin. The normalized gene densities per deletion bin were calculated by dividing the density for each bin by the mean of the densities along chromosome 3B. The Rice Genome Annotation Project from Michigan State University (http://rice.plantbiology.msu.edu/, Rice Pseudomolecules v6.1 database) was used to estimate the number of annotated genes on the rice chromosomes. For the Chi² tests, to test the uniformity of the percentages and the densities along chromosome 3B, we estimated the number of genes per deletion bins that would have generated a uniform distribution and we used these numbers as theoretical values.

Authors' contributions

PH and JM carried out the hybridisations; CR, PH and FC performed data analysis; RW, CF and EP helped with the interpretation of the results; CR drafted the manuscript; FC, RW, CF and EP were involved in improving the manuscript. The final version of the manuscript was approved by all the authors.

Abbreviations

MTP: minimal tiling path; TE: transposable element.

Acknowledgements

The authors would like to thank Christel Laugier for interesting discussions about data analysis. The research leading to these results has received funding from the European

Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement n°FP7-212019. CR was financially supported by Région Auvergne.

Conclusion Article n°2

L'hybridation de pools de BAC sur une puce portant des unigènes d'orge a permis de cartographier plus de 700 gènes sur les BAC du chromosome 3B de blé tendre de façon efficace et précise. L'analyse de leur distribution a permis de mettre en évidence un gradient positif de densité du centromère vers les télomères du chromosome 3B. De plus, 40% des gènes ont été positionnés à proximité d'au moins un autre gène. Ces gènes ont été considérés comme faisant partie d'îlots de gènes potentiels alors que les 60% restant ont été considérés comme isolés. Enfin la distribution des gènes en îlots le long du chromosome suggérait que la densité d'îlots était responsable du gradient de la densité de gènes observé.

Cependant, cette analyse n'a permis d'appréhender qu'une fraction limitée des gènes portés par le chromosome 3B. En effet, la puce ne portait que 15000 unigènes alors que Choulet et al. (2010) ont estimé que chaque sous-génome de blé devait compter environ 50000 gènes. De plus, l'utilisation de cette puce d'orge a potentiellement limité l'efficacité de l'approche du fait de la divergence de séquence entre le blé et l'orge.

Pour tirer le plein parti de cette approche, nous avons développé une puce portant des unigènes de blé pour améliorer la performance de l'expérience. Combinée avec une nouvelle version de la carte physique couvrant désormais la quasi-totalité du chromosome 3B, cette puce a conduit à l'établissement de la première carte transcriptionnelle d'un chromosome de blé. Cette carte nous a permis d'étudier l'organisation, la régulation et l'évolution de l'espace génique du blé avec une exhaustivité et une résolution jamais atteintes à l'heure actuelle.

ARTICLE N°3

La carte transcriptionnelle du chromosome 3B portant 3000 gènes a révélé des tendances spécifiques sur l'organisation de l'espace génique et la régulation chez le blé hexaploïde

A 3000-loci transcription map of the chromosome 3B unraveled specific patterns of gene space organization and regulation in hexaploid wheat

Camille Rustenholz[1], Frédéric Choulet[1], Christel Laugier[1], Jan Safar[2], Hana Simkova[2], Jaroslav Dolezel[2], Federica Magni[3], Simone Scalabrin[3], Federica Cattonaro[3], Sonia Vautrin[4], Arnaud Bellec[4], Hélène Bergès[4], Catherine Feuillet[1] and Etienne Paux[1*]

[1] *INRA UMR 1095, Génétique Diversité et Ecophysiologie des Céréales, F-63100 Clermont-Ferrand, France*

[2] *Laboratory of Molecular Cytogenetics and Cytometry, IEB, CZ-77200 Olomouc, Czech Republic*.

[3] *Istituto di Genomica Applicata, Parco Scientifico e Tecnologico di Udine "Luigi Danieli", I-33100 Udine, Italy*

[4] *INRA-Centre National de Ressources Génomiques Végétales, F-31326 Castanet Tolosan, France*

[*]Corresponding author

ABSTRACT

To improve our understanding of the organization and regulation of the wheat gene space, we established the first transcription map of a wheat chromosome (3B) by hybridizing a newly developed wheat expression microarray with BAC pools from a new version of the 3B physical map as well as with cDNA probes from five tissues at three developmental stages each. Mapping data for almost 3000 genes on 3B showed that the gene space spans the whole chromosome with a twofold increase of gene density towards the telomeres due to an increase of genes in islands. Moreover comparative genomics with rice and *Brachypodium* revealed that these gene islands are mainly composed of non-collinear genes. Such intra- or interchromosomal gene duplications might originate from a combination of factors including transposable element (TE) preferential insertion in particular parts of the genome, TE-mediated gene capture or repair of double strand breaks due to TE insertions or recombination events by gene-containing foreign fragments. In addition, gene ontology and expression profiles along the chromosome revealed that these islands are enriched in genes sharing the same function or expression profiles suggesting that genes in islands have acquired specific regulations during the evolution. Finally, co-regulation islands have been identified demonstrating the existence of long-distance regulation mechanisms in wheat.

INTRODUCTION

The organization of the gene space in a genome refers to the layout of the protein-coding genes along the chromosomes. With the growing number of sequenced genomes, many studies led to the conclusion that this organization is far from random and is correlated to the genome size. For example plants with the smallest genomes such as *Arabidopsis thaliana* (125 Mb), *Brachypodium distachyon* (272 Mb) and *Oryza sativa* (389 Mb) exhibit an even distribution of their genes along their genome (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; The International Brachypodium Initiative, 2010) whereas for the intermediate size genomes of *Populus trichocarpa* (485 Mb) and *Vitis vinifera* (487 Mb), alternations between high gene density regions and low gene density regions are observed (Tuskan et al., 2006; Jaillon et al., 2007). This tendency is even stronger in the large plant genomes of *Glycine max* (1115 Mb) and *Zea mays* (2300 Mb) in

which a positive gradient of the gene density from the centromere to the telomeres was observed (Schnable et al., 2009; Schmutz et al., 2010). The only permanent feature that is found among all the eukaryotic genomes is that the gene density drops dramatically nearby the centromeres of the chromosomes (Lomiento et al., 2008).

The genome of bread wheat, *Triticum aestivum* L., is one of the most complex plant genomes as it is allohexaploid (A, B and D genomes), counts 17000 Mb of sequence (~5700 Mb per subgenomes) and more than 80% of repeated sequences. As a consequence, molecular analyses on the wheat genome have always been very challenging. Therefore the wheat genome sequence is still not available and little is known about the organization of the gene space along the wheat chromosomes. Some evidences of an uneven distribution of the genes along the wheat chromosomes were given by Devos et al. (2005) and Charles et al. (2008) who sequenced randomly chosen BAC clones. Indeed they found BACs carrying genes and BACs without genes as it could be expected from large plant genomes. Moreover Munkvold et al. (2004) based on EST mapping in wheat deletion bins and Choulet et al. (2010) based on the annotation of megabase-sized sequences suggest a positive gradient of the gene density from the centromere to the telomeres.

Within the *Triticum* lineage, it has been previously suggested that the gene space was organized in isolated genes and gene islands, also called gene-rich regions, of two or more genes located close to each other (Brooks et al., 2002; Chantret et al., 2004; Wicker et al., 2005; Choulet et al., 2010; Rustenholz et al., 2010). Choulet et al. (2010) identified that 50% of the wheat intergenic distances were shorter than 43 kb and used this distance to define a gene island in the wheat genome. Based on this definition, islands of two to ten genes with three genes on average were identified. Moreover, the distribution of gene islands was shown to strongly explain the positive gradient of the gene density along the wheat chromosome 3B (Rustenholz et al., 2010). Such gene islands have also been described in other genomes, such as maize where Wei et al. (2009) identified that 56% of the intergenic distances were shorter than 20 kb which is comparable to what was found in the wheat genome after genome size correction. Comparable gene densities for gene islands were also found in the soybean and in the cotton genomes (Clough et al., 2004; Guo et al., 2008). However so far little is known about the formation of the gene islands and about the possible forces than maintain some genes close to each other in islands.

In this study, we mapped physically almost 3000 genes, defined as transcriptional units, on wheat chromosome 3B BACs using a newly developed wheat NimbleGen 40K unigene microarray and the new version of the wheat chromosome 3B physical map. Simultaneously we also generated their expression profiles from various wheat organs at different developmental stages to establish the first transcription map of a wheat chromosome. We confirmed that 70% of the 3000 chromosome 3B genes are organized in gene islands and are responsible for the positive gradient of gene density from the centromere to the telomeres. By studying their evolution, their expression and their function, we found co-expression and co-function gene islands among which some were conserved with rice and/or *Brachypodium* and some were newly formed in the wheat genome. We finally suggested structural and functional hypotheses for the formation and the conservation of the genes organized in islands.

RESULTS

A sequence-ready physical map of the chromosome 3B of hexaploid wheat

In 2008, we published the first physical map of a wheat chromosome, 3B (995 Mb) (Paux et al., 2008). This was a major breakthrough in wheat genomics as it demonstrated the potential of the chromosome-specific approach to tackle the wheat genome, opening the way to future genome sequencing and accelerated improvement of this essential crop species. However, this map covered only 82% of the chromosome, which was not compatible with a whole chromosome sequencing project. To circumvent this limitation, we thus decided to extend this physical map, moving from a chromosome landing-ready to a sequencing-ready map. To this aim, a new chromosome 3B-specific BAC library was constructed containing 82,176 BAC clones. These clones together with the 7440 clones from the minimal tiling path (MTP) of the first version were fingerprinted and the resulting high information content fingerprints (HICFs) were added to the 1036 contigs from the first version of the physical map of the wheat chromosome 3B previously published by Paux et al. (2008). The assembly of 131,792 HICFs using FPC (Soderlund et al., 1997; Soderlund et al., 2000) resulted in 1669 contigs covering 961 Mb (97% of the whole chromosome) with a 19.2 X coverage depth. Using mapping information from the 1443 markers already assigned to contigs (Paux et al., 2008), 919 out of the 1669 contigs covering 740 Mb were assigned to one of the eight intervals

Table 4 : Number and density of loci, number of rearranged genes and number of coexpressed genes per deletion bin of wheat chromosome 3B.

3B Deletion Bin	Contig size assigned per bin (Mb)	Number of loci	Locus density (locus / Mb)	Number of rearranged genes	Number of coexpressed genes
3BS8-0.78-1.00	55.7	205	3.68	178	16
3BS9-0.57-0.78	72.9	193	2.65	155	12
3BS1-0.33-0.57	124.4	354	2.85	287	22
C-3BS1-0.33	74.2	184	2.48	153	10
C-3BL2-0.22	56.5	159	2.81	122	9
3BL2-0.22-0.50	103.9	300	2.89	232	14
3BL10-0.50-0.63	46.5	140	3.01	101	4
3BL7-0.63-1.00	206.4	661	3.20	521	60
Total assigned	**740.7**	**2196**	**2.96**	**1749**	**147**
Not assigned	220.3	728	/	637	39
Total	961	2924	/	2386	186

defined by genetic deletions, so-called "deletion bins" (3BS8-0.78-1.00, 3BS9-0.57-0.78, 3BS1-0.33-0.57, C-3BS1-0.33, C-3BL2-0.22, 3BL2-0.22-0.50, 3BL10-0.50-0.63 and 3BL7-0.63-1.00) (Table 4). A subset of 9216 BACs representing the minimal tiling path (MTP) was selected and rearrayed. These clones are currently being used to sequence the wheat chromosome 3B using a BAC-by-BAC approach based on second generation sequencing. In parallel, 64 three-dimensional (plate, row, and column) MTP pools were produced and subsequently used in this study.

A chromosome-wide survey of the wheat gene space organization

Hybridizing expression microarray with BAC pools has already been demonstrated to be a very efficient technique to assign individual genes to physical maps (Rustenholz et al., 2010). To this aim, a wheat NimbleGen 40K unigene microarray was developed based on the *Triticum aestivum* NCBI unigene build #55 (February 2009; http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=4565). It contains on average three 60-mer probes for 39,179 unigenes out of the 40,349 of the complete set.

This chip was hybridized with the 64 MTP pools of the chromosome 3B physical map. After signal quantification, normalization and data deconvolution, 2913 unigenes were unambiguously assigned to 3003 loci, 78 unigenes being located at multiple positions on chromosome 3B (two to four positions). Out of these 2913, 2142 genes (71%) were found in the vicinity (same BAC or overlapping BACs) of at least another gene. To assess whether these neighbor genes corresponded to one single locus or two independent loci, we rebuilt the corresponding EST contigs using NCBI EST dataset. We then performed pairwise sequence comparisons of the resulting unigenes as well as a homology search against the rice and *Brachypodium* genomes. In total, 159 neighbor genes (corresponding to 69 groups) showing at least 98% sequence identity and matching potentially the same gene were identified. After removing genes that were considered as redundant, 2836 unigenes (2924 gene loci) were unambiguously located on wheat chromosome 3B physical map and mapped to 2071 BACs belonging to 1016 individual contigs of 69 kb to 3.4 Mb.

To assess the reliability of these results, hybridization data on 13 sequenced contigs were compared to sequence annotations (Choulet et al., 2010). Out of the 73 unigenes mapped to

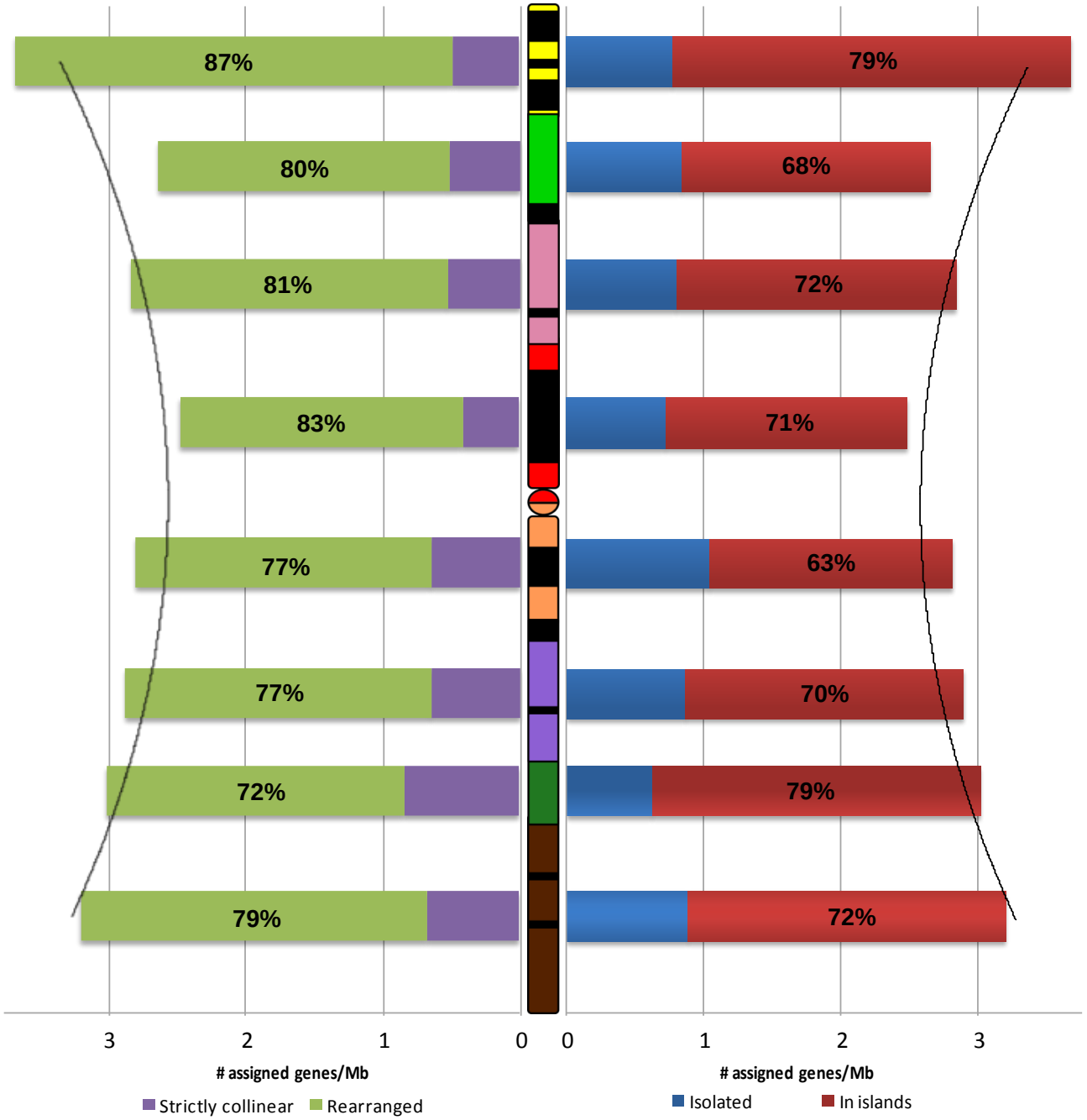


Figure 27 : Gene distribution along the wheat chromosome 3B. Each colour on wheat chromosome 3B corresponds to a deletion bin. Yellow: 3BS8-0.78-1.00 ; light green: 3BS9-0.57-0.78 ; pink: 3BS1-0.33-0.57 ; red: C-3BS1-0.33 ; orange: C-3BL2-0.22 ; purple: 3BL2-0.22-0.50 ; dark green: 3BL10-0.50-0.63 and brown: 3BL7-0.63-1.00. The black segments correspond to the heterochromatic regions identified by C-banding, the coloured segments to the euchromatic regions and the circle to the centromere. On the left, the density of strictly collinear genes is represented by purple bars. The density of the rearranged genes is represented by green bars. The proportions of rearranged genes per deletion bins are shown as percentages within the green bars. On the right, the density of isolated genes is represented by blue bars. The density of the genes organized in island is represented by red bars. The proportions of genes organized in island per deletion bins are shown as percentages within the red bars. The two black curves represent the regression curve of the gene density.

the corresponding contigs, 52 (71%) were confirmed by BLAST analysis. Thirty one (60%) matched a region previously annotated as gene (or pseudogene), 17 (33%) matched a non annotated region and four (8%) matched a transposable-element associated region. The remainder (21 unigenes) was assigned to these contigs but their position was not supported by sequence information. For four of them, one single probe out of the three showed positive hybridization signals and can therefore be considered as false positives. Ten (48%) were clustered in groups of two to three genes on the same BAC. It is very unlikely that false positive clones could have been detected by several independent probes and strongly supports the mapping position of the corresponding unigenes. Such discrepancy between hybridization and sequencing data might then originate from inconsistencies in contig assembly between the first version of the physical map used for sequencing and the second one used in hybridization experiments. For the remaining 7 unigenes, no clear trend was identified. Several hypotheses can be proposed, including false positives, misassembled MTP BACs or gaps in the sequence. However, despite these 21 genes, these data validated this method as a powerful and reliable approach to map genes to BAC contigs and investigate the gene space organization.

As previously stated, the new version of the physical map counts 1669 contigs, out of which 919 are assigned to one of the eight chromosome 3B deletion bins. These results were used to map 2196 genes (75%) to one of these deletion bins and therefore calculate the gene density for each bin using the cumulated length of anchored contigs from the physical map in each deletion bin (Table 4). We found that the gene density distribution was significantly positively correlated with the distance to the centromere (Pearson's correlation coefficient r=0.717; *P*-value=0.045), suggesting a gradient of gene density from the centromere to the telomeres (Figure 27). The distal 3BS8-0.78-1.00 deletion bin showed the highest gene density with 3.7 genes per megabase and the proximal C-3BS1-0.33 deletion bin the lowest with 2.5 genes per megabase. In addition the gene density along chromosome 3B was significantly positively correlated to crossing-over frequency (Pearson's correlation coefficient r=0.719; *P*-value=0.044). One limitation of the hybridization-based gene mapping relies in the fact that it cannot discriminate between one or more copies of the same gene on the same BAC. However, Choulet et al. (2010) estimated that the tandemly duplicated genes are mainly located at the distal parts of the chromosome where they represent ~30% of the genic content. Based on this assumption and on the estimation of 6000 genes on chromosome 3B without the tandemly duplicated genes (Paux et al., 2006), we extrapolated the gene density of the distal

3BS8-0.78-1.00 to 9.8 genes per megabase (1 gene every 102 kb) and that of the proximal C-3BS1-0.33 deletion bin to 5.1 genes per megabase (1 gene every 197 kb). Therefore when considering the tandemly duplicated genes, the telomeric part of the short arm of wheat chromosome 3B showed a twofold increase (1.9) in gene density compared to the centromeric part. This gene density distribution slightly differs from the one we estimated previously using a barley Agilent 15K unigene microarray (Rustenholz et al., 2010), especially in the most distal deletion bin of the short arm, 3BS8-0.78-1.00. However, only 30% of the wheat microarray unigenes are present on the barley microarray and this may impact the reliability of the experiment. In addition, this proportion varies between deletion bins, ranging from 26% for 3BS8-0.78-1.00 to 34% for 3BL10-0.5-0.63. These variations were shown to explain the differences in the gene density distribution between wheat and barley (Pearson's correlation coefficient r=-0.733; *P*-value=0.039). Indeed the more wheat unigenes are absent from the barley set, the more the gene densities established with the wheat and the barley data are different per bin. Thus the densities established with the wheat microarray seem more reliable.

We then assessed the conservation level between wheat and rice or *Brachypodium*. Out of the 2836 wheat chromosome 3B unigenes, 1519 (54%) and 1564 (55%) orthologous genes were found in rice and *Brachypodium*, respectively. Out of them, 877 (58%) and 1016 (65%) unigenes were syntenic, *i.e.* orthologous to genes located on the orthologous rice chromosome 1 and in *Brachypodium* chromosome 2, respectively regardless of the gene order (Keller and Feuillet, 2000). We then studied the collinearity, *i.e.* the gene order (Keller and Feuillet, 2000), between wheat and the two model grass species. We found 586 collinear genes which represented 44% of the genes from contigs containing at least two rice or *Brachypodium* orthologs. Interestingly, 8% of these collinear genes were not in syntenic position on rice chromosome 1 or *Brachypodium* chromosome 2, suggesting wheat-specific rearrangements involving blocks of genes.

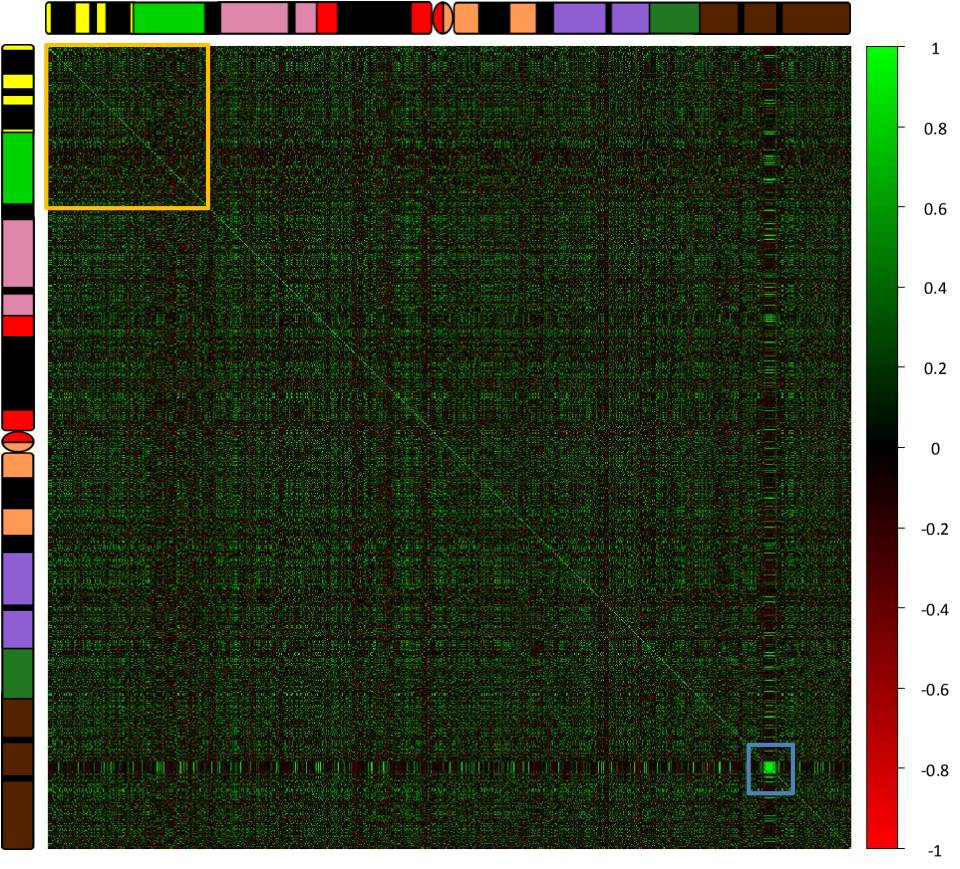
By comparing the distribution of collinear syntenic genes (hereafter referred to as 'strictly collinear genes') with the distribution of other genes (*i.e.* non syntenic genes, non collinear genes and genes with no orthologs in rice and *Brachypodium*) (hereafter referred to as 'rearranged genes'), we found that the gradient of gene density along chromosome 3B was mainly due to the presence of rearranged genes (Pearson's correlation coefficient r=0.934; *P*-value=7E-4) whereas strictly collinear genes have no impact on the overall gradient (Pearson's correlation coefficient r=0.162; *P*-value=0.702) (Table 4 and Figure 27). In

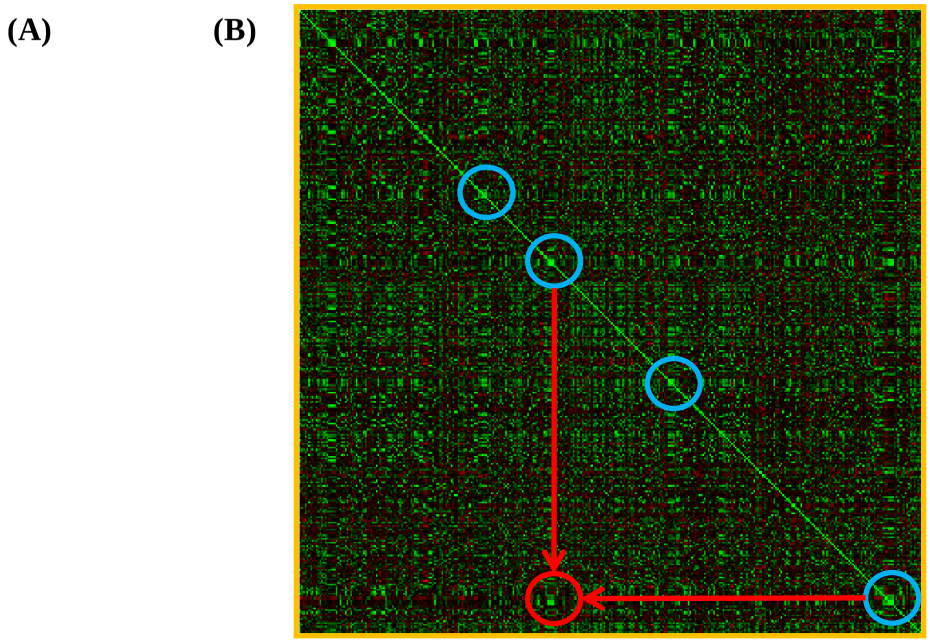
addition, the density of rearranged genes was found to be correlated to the crossing-over rate (Pearson's correlation coefficient r=0.727; *P*-value=0.041).

To further investigate the gene space organization, we defined gene islands as genes located on the same or on overlapping BACs. We found 2040 genes (70%) being part of 709 gene islands composed of two to 30 genes (average=2.9 ± 1.6, median=2). This proportion of genes in island is far different from random as 10,000 random samplings without replacement of 2924 gene locations on the chromosome 3B BACs never reached this percentage (average=57.9 ± 1.0%, *P*-value=0). The remaining 884 genes (30%) were considered as isolated. The density of isolated genes was shown to be uniform along chromosome 3B and not correlated to the distribution of the total gene density (Chi² test, *P*-value=0.347; Pearson's correlation coefficient r=-0.077, *P*-value=0.857). By contrast, the density of genes in islands varies among deletion bins (ranging from 63% in C-3BL2-0.22 to 79% in 3BL10-0.50-0.63) and is positively correlated to the distribution of the total gene density (Pearson's correlation coefficient r=0.951, *P*-value=3E-4) (Figure 27). Moreover, this density was also correlated with the density of rearranged genes (Pearson's correlation coefficient r=0.885; *P*-value=0.003) whereas no correlation was found with the density of strictly collinear genes (Pearson's correlation coefficient r=0.515; *P*-value=0.192). These results demonstrate that the gradient of gene density observed along chromosome 3B results from an increase of genes in islands from the centromere to the telomeres and that these gene islands mainly originate from gene rearrangements and not from genes that were already close in an ancestral genome.

Transcription mapping of the chromosome 3B

To study the relationships between gene space organization and gene expression, we established the transcription map of chromosome 3B by hybridizing 15 cDNA samples originating from five wheat organs (root, leaf, stem, spike and grain) at three developmental stages each onto the wheat NimbleGen 40K unigene microarray. After signal quantification and data normalization, transcript profiles were drawn for 32,284 unigenes (82%) including 2515 of the 2836 (89%) located on chromosome 3B. The remaining 6895 unigenes (321 from 3B) did not show significant hybridization signals in any of the 15 samples.





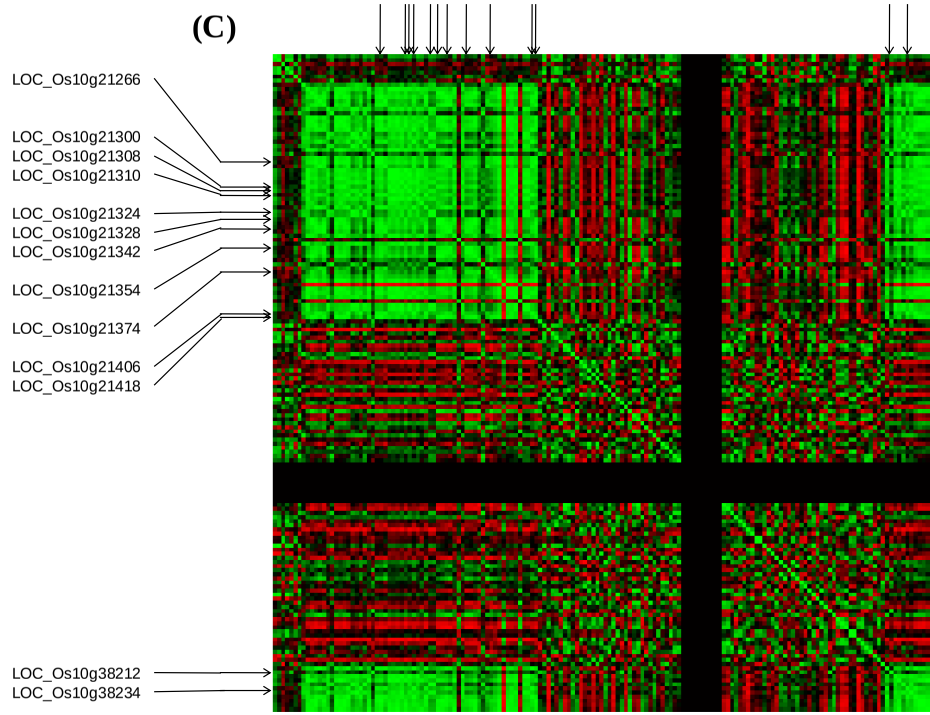


Figure 28 : Correlation maps of gene expression profiles. The correlations of gene expression profiles are represented using the green-red scale shown. Light green corresponds to positive correlation and light red to anti-correlation. (A) Correlation map of chromosome 3B. The chromosome 3B is depicted as previously. The contigs were ordered as described in the "Methods" section. The green clusters on the diagonal line correspond to coexpression clusters with the largest squared in blue. (B) Zoom of the orange squared region of (A). The coexpression clusters are circled in blue. A coregulation island is circled in red. (C) Correlation map of expression profiles of rice genes on chromosome 10. The map goes from LOC_Os10g21100 to LOC_Os10g38278. A gap (black regions) was created for a better representation. The orthologous of wheat coexpressed genes are indicated with arrows with their names on the left of the map.

The expression profiles were validated through quantitative reverse-transcription PCR (QRT-PCR) on eight genes. The QRT-PCR expression profiles of seven genes were highly correlated with the profiles established through hybridization (Pearson's correlation coefficient r>0.7), whereas for the last gene, no amplification was achieved, probably due to technical problems in primer design.

After hierarchical clustering of the unigenes located on chromosome 3B with a Pearson's correlation coefficient threshold of 0.641 (*P*-value=0.01), 153 groups of unigenes showing similar expression profiles in the 15 samples were established. Out of the 1727 "expressed" genes in islands, 186 (11%) were co-expressed with their neighbor(s), defining 80 so-called 'co-expression clusters' of two to 21 genes (average=2.3 ± 2.1, median=2) (Figure 28A and B). This proportion (11%) is far different from random as 10,000 random samplings without replacement of the gene locations on the BACs never reached this percentage (average=3.9 ± 0.6%, *P*-value=0). In addition, the density of co-expressed genes in islands was significantly not uniform along chromosome 3B (Chi² test, *P*-value=0.009) with more co-expressed genes in islands than expected in the two most telomeric deletion bins (13% in 3BS8-0.78-1.00 and 15% 3BL7-0.63-1.00) (Table 4).

In an attempt to assign putative functions to the wheat 3B unigenes, we used the Gene Ontology (GO) of rice and *Brachypodium* orthologs. GO terms (*i.e.* biological process, function and cellular component location) were associated to 1277 unigenes out of the 2836 unigenes located on wheat chromosome 3B (45%) and 657 genes with a GO term were located in gene islands. To look for differential distribution of gene functions along chromosome 3B, we calculated the proportion of each GO term in each deletion bin. Three GO terms displayed a significantly non-uniform distribution along the chromosome: binding (GO:0005488) (Chi² test, *P*-value=0.025), transporter activity (GO:0005215) (Chi² test, *P*-value=0.034) and plastid (GO:0009536) (Chi² test, *P*-value=0.011). The distribution of the transporter activity GO term along chromosome 3B was negatively correlated with the distance to centromere and the total gene density (Pearson's correlation coefficient r=-0.832 and -0.941, *P*-value=0.010 and 5E-4, respectively) whereas the two other GO term distribution did not correlate with any variable. Out of the 657 genes in island associated to a GO term, 370 (56%) located in 153 gene islands shared at least one GO term with their neighbor(s). This result is far different from random as demonstrated by 10,000 random samplings without replacement of the gene locations on BACs (average=47.3 ±2.3%, *P*-

value=0), strongly suggesting that genes involved in the same biological process or sharing the same function or cellular localization tend to occur at adjacent or nearby positions in the genome. Furthermore, out of the 63 co-expressed genes in islands with a GO term, 92% (58/63) also shared one or more GO term(s) with their neighbor(s). This result is different from random as 10,000 random samplings without replacement of the gene locations on the BACs rarely reached this percentage (average=36.9 ± 19.0%, *P*-value=0.009). Thus the co-expression clusters that we identified on wheat chromosome 3B are significantly enriched in genes sharing the same GO term(s).

We further investigated the conservation of the co-expressed clusters between wheat and rice or *Brachypodium*. Out of the 80 co-expression clusters identified on chromosome 3B, 54 (68%) were composed of non collinear genes, *i.e.* genes not close to each other in the rice and/or in the *Brachypodium* genome. Eighteen co-expression clusters (23%) were found to be composed of strictly collinear genes. A detailed analysis of the genes present in these 18 clusters revealed that 89% showed identity to the same rice and/or *Brachypodium* gene and were very likely corresponding to old tandemly duplicated genes (as redundant unigenes corresponding to the same gene have been previously deleted). Interestingly, we also identified eight clusters of non syntenic collinear genes, probably originating from genomic rearrangements of blocks of genes. A striking example is a gene island located on the distal bin of the long arm of chromosome 3B (3BL7-0.63-1.00). It contains 30 genes on 474 kb which represents a gene density of 1 gene every 16 kb. Out of the 30 genes, only one did not share any functional feature (expression or GO term) with another member of the island. For the 29 other genes, eight were co-expressed, one shared the same GO terms with other members of the island and 20 were both co-expressed and shared the same GO term(s). Thus this cluster of genes is highly co-expressed and co-functional. Out of the 30 genes, 22 genes were non-syntenic genes and mainly orthologous to genes located on rice chromosome 10 and *Brachypodium* chromosomes 3 and 5. Most notably, a block of 16 collinear genes orthologous to the rice chromosome 10 genes was identified. This region contains mainly house-keeping genes such as genes coding for chlorophyl-associated proteins, ribosomal proteins or phytochrome. Out of these 16 genes, 12 were co-expressed and co-functional, three were co-expressed only and one did not share any functional feature with another member of the group. Moreover, using transcriptomic data of rice chromosome 10 (see Methods), the orthologous genes in rice were found highly co-expressed as well (Figure 28C). Therefore

despite a large genome rearrangement that led to a synteny break, collinearity as well as co-regulation were maintained in wheat and rice during the evolution. Surprisingly, this cluster is not conserved in *Brachypodium* as most of the genes were not found in this genome and the remainder was located on three different chromosomes (namely 1, 3 and 5).

Furthermore, the transcription correlation map of chromosome 3B (Figure 28B) also revealed correlations between the expression patterns of distant co-expression clusters. Eighteen groups, so-called "co-regulation islands", were found on chromosome 3B and were composed of two to seven distant co-expression clusters showing the same expression profiles. These co-regulation islands involved 54 out of the 80 co-expression clusters (68%) and 132 out of the 186 co-expressed genes (71%). In addition to their expression profiles, most of the genes involved in these co-regulation islands share the same GO terms. Moreover the density of the genes involved in co-regulation islands was significantly positively correlated to the distance to the centromere (Pearson's correlation coefficient r=0.730, *P*-value=0.040). For example the telomeric 3BS8-0.78-1.00 deletion bin showed 100% of co-expressed genes involved in a co-regulation island whereas the centromeric C-3BS1-0.33 deletion bin showed 40%. In terms of synteny, eight co-regulation islands involved only rearranged genes, nine involved both rearranged and strictly collinear genes and only one involved only strictly collinear genes. Thus the co-regulation islands on the same chromosome seem to be marginally conserved between wheat and rice or *Brachypodium*.

DISCUSSION

Transcription mapping in wheat has the potential to support genome annotation through the identification of unannotated transcriptional units

By hybridizing a newly developed wheat NimbleGen 40K unigene microarray with MTP BAC pools from the new version of the chromosome 3B physical map as well as with 15 cDNA samples originating from five different tissues at three developmental stages each, we established the first transcription map of a wheat chromosome. This map contains 2924 gene loci corresponding to 2836 unigenes, 2515 of which being associated with transcript profiles and 1277 with GO terms.

Comparison of our mapping data with the annotation of 13 sequenced contigs on chromosome 3B (Choulet et al., 2010) confirmed the hybridization-based mapping positions for 71% of the unigenes. For the remaining 29%, mapping positions were not confirmed by sequence data. However, for half of them, the clustering of two to three genes per BAC strongly supports the validity of their mapping position, therefore suggesting local misassembly regions in the physical map. Indeed the MTP BACs used to sequence contigs did not necessarily correspond to the MTP BACs used for hybridizations as they originated from the two successive versions of the physical map. Such misassembly points are not unexpected as previous experiments led to an estimate of ~10% of misassembled BACs in the chromosome 3B physical map (unpublished data). All in all, this means that 85% of the mapping data should be correct and confirms that unigene microarray hybridization is a powerful and reliable technique as demonstrated previously using Agilent 15K chip (Rustenholz et al., 2010). Interestingly, this comparison also unraveled 17 unigenes that were not annotated so far. For 70% of them, a significant expression was observed in at least one of the 15 cDNA samples, confirming that they do correspond to transcriptional units. Such identification by tiling microarrays of new transcriptional units that had not been identified through computational annotation have already been reported in other species such as *Escherichia coli*, *Arabidopsis thaliana*, fruit fly, human and rice (Bertone et al., 2005; Jiao et al., 2005; Stolc et al., 2005; Gregory et al., 2008). Some of these transcriptional units show the classical structural features of a coding gene (Stolc et al., 2005) but some others can be duplicated gene fragments, antisense transcripts or potential non-coding RNAs (Bertone et al., 2005; Li et al., 2007). Bertone et al. (2004; 2005) noticed that a large proportion of the new transcriptional units they identified where conserved between the human and other mammalian genomes at the protein level. Thus they suggested that these new transcriptional units may be functional on the basis of evolutionary conservation. A detailed analysis of these 17 new transcriptional units identified on chromosome 3B is currently underway.

Genome rearrangements shaped the wheat genome through the formation of gene islands

By mapping almost 3000 gene loci along wheat chromosome 3B, this transcription map allowed us to perform the most comprehensive analysis of the wheat gene space to date. This latter has long been a matter of controversy. In 2004, Erayman et al. (2004) suggested that

most of the genes were clustered in small genomic regions located in the distal parts of the wheat chromosomes. Devos et al. (2005) and Charles et al. (2008) reached different conclusions and suggested that the gene density was one gene per BAC on average based on random BAC sequencing. More recently, Choulet et al. (2010) showed that the gene space was spanning the whole wheat chromosome 3B without megabase-sized geneless regions.

Our transcription map confirmed this finding and also revealed an increasing gradient of gene density from centromere to telomeres. Based on the mapping of almost 3000 genes and assuming a number of ~8000 genes with 30% of tandem duplications on chromosome 3B (Choulet et al., 2010), we extrapolated that the gene density in the most telomeric part of the wheat chromosome 3B short arm should be one gene every 102 kb and one gene every 197 kb in the most centromeric part. This twofold gradient of the gene density between the centromere and the telomeres was also reported by Munkvold et al. (2004) and more recently by Choulet et al. (2010) who found one gene every 86 kb in the more distal regions and one gene very 184 kb in the proximal regions. Following a similar approach on barley 15K expression chips, we recently suggested that this gradient of gene density might be explained by a non-uniform distribution of genes in islands (Rustenholz et al., 2010). However, in this study, because of the limited number of genes mapped on chromosome 3B (757), only 40% of the genes were found to be part of gene islands. Here, with almost 3000 gene loci, we refined this proportion to 70%, with islands composed of two to three genes. These results are consistent with previous estimates based on megabase-sized contig sequences (Choulet et al., 2010). In addition, we also observed the differential distribution of gene in islands, ranging from 63% in proximal bins to 79% in distal bins and confirmed the strong correlation between the overall gene density gradient and the proportion of genes in islands.

Comparative analyses between wheat chromosome 3B and rice chromosome 1 genes suggested an overall synteny level of 58%, which is consistent with previous studies (La Rota and Sorrells, 2004; Munkvold et al., 2004; Varshney et al., 2005; Bilgic et al., 2007; Stein et al., 2007a). The synteny level was slightly higher with *Brachypodium* chromosome 2 (65%), as expected based on the divergence time between these two species, as well as on other studies (Bossolini et al., 2007; The International *Brachypodium* Initiative, 2010). In addition we found a collinearity level of 44% which is slightly lower than what Choulet et al. (2010) found in their study, *i.e.* 52%. This discrepancy might be due to the fact that we mapped only one third of the expected genes along chromosome 3B. Therefore if one member of a pair of

collinear genes was missing, we could not consider the remaining gene as collinear. However, neither the synteny nor the collinearity level was even along the chromosome 3B. Detailed analysis in each deletion bin revealed a negative gradient of conservation towards the telomeres. In addition, it also revealed a clear negative correlation between the conservation level and the proportion of genes in islands. All in all, our data strongly suggest that the gradient of gene density along the chromosome is due to an increase in the proportion of genes in islands and that most of these islands originate from genome rearrangements, rather than from genes that were already close to each other in an ancestral genome and maintain during the evolution.

Gene islands, also called gene-rich regions, are common features of large and highly repetitive plant genomes, especially at the distal parts of chromosomes (Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). Different evolutionary scenarios have been proposed to explain the formation of these islands. Among them, "passive" mechanisms have been suggested that do not imply gene movement but rather differential insertion or deletion of TEs into the genome to create gene islands from an ancestral backbone. A first example of these "passive" mechanisms was suggested for the sorghum where homogeneous expansion of the genome combined with preferential deletions in gene-rich regions would lead to the gene space distribution observed in this genome (Paterson et al., 2009). This scenario was invalidated for wheat by Choulet et al. (2010) who found that solo-LTRs that are marks of LTR retrotransposon deletion through recombination were mainly found in large gene-poor regions. A second "passive" mechanism was proposed for different species including maize in which TEs insert preferentially in particular parts of the genome depending on their integrase affinity with specific chromatin protein (Baucom et al., 2009a). Such preferential insertion pattern was also observed in wheat where LTR retrotransposons such as *gypsy* and *copia* display a clear insertion bias into each other (Paux et al., 2006) and therefore might be responsible for gene islands involving collinear genes.

However, we demonstrated that a significant part of gene islands are composed of "rearranged" genes, therefore involving more "active" mechanisms where genes are duplicated, translocated into the genome and added to the ancestral gene backbone to create gene islands. Such mechanism has been proposed for wheat by Choulet et al. (2010). Indeed, these authors found two different evidences suggesting TE-mediated interchromosomal gene duplications. The first one was a significant correlation between the proportion of rearranged

genes and the age of TEs, as an estimate of their activity: the younger the TEs, the lower the synteny level. The second evidence came from the detection in chromosome 3B sequences of four genes fully included into CACTA transposons. However, if TE-driven gene capture is very likely to be responsible for gene movement in wheat, it cannot explain alone the high proportion of rearranged genes. Indeed, based on the four genes discovered in chromosome 3B sequences, Choulet et al. (2010) estimated that CACTAs would be involved potentially in capturing ~200 gene fragments on chromosome 3B, the same number reported for the sorghum genome (Paterson et al., 2009). Even if other transposable elements can be involved in gene amplification and mobilization in wheat, it is very unlikely that this mechanism can have led to the movement of almost 4000 genes, as estimated by the overall synteny level of 44-52% between wheat and rice. Wicker et al. (2010) reached the same conclusion in a recent study where they performed a three-way comparison of *Brachypodium*, rice, and sorghum genomes. In addition to the TE-driven gene capture, the authors proposed a mechanism where the foreign fragments containing genes were introduced as ‘‘filler DNA’’ to repair double-strand breaks (DSBs) that occurred upon TE insertion or a recombination event. Interestingly, our study clearly showed a strong correlation between the proportion of rearranged genes and the crossing-over rate, suggesting a putative relationship between gene movement and recombination.

Thus we conclude that gene islands in wheat originate from a combination of different factors including the preferential insertion of LTR retrotransposons into each other to create “TE oceans”, the TE-mediated gene movement and the repairing of DSBs produced by TEs or recombination. However the relative contribution of these different factors remains to be determined.

Gene order is not random in the wheat genome and co-expressed and co-functional genes tend to be linked

Beside mechanistic aspects of gene island formation, functional reasons such as gene regulation and co-expression have been proposed as driving forces for keeping genes close to each other in islands (Hurst et al., 2004; Batada et al., 2007; Babu et al., 2008; Chen et al., 2010a; Choulet et al., 2010). Numerous studies detailed the two main mechanisms involved in the *cis*-regulation of these adjacent co-expressed genes. Firstly promoters either bidirectional

or not, enhancers, regulatory sequences binding transcription factors can induce co-expression of multiple genes from short distances (Hurst et al., 2004; Babu et al., 2008; Chen et al., 2010a). Such regulatory elements could lead to polycistronic transcripts, *i.e.* transcripts containing two or more open reading frames, implying perfect co-expression as observed in *Arabidopsis* (Thimmapuram et al., 2005; Chen et al., 2010a). However, Chen et al. (2010a) suggested that such direct interactions are enhanced at gene distances below 400 bp in *Arabidopsis*. For large intergenic distances, chromatin modifications would be a favored mechanism to induce co-expression of multiple genes as they allow regions to switch between suppression (heterochromatin) and activation of transcription (euchromatin) (Hurst et al., 2004; Batada et al., 2007; Babu et al., 2008; Chen et al., 2010a).

In the human, mouse, *Arabidopsis* and rice genomes, the percentage of adjacent co-expressed genes ranges from 2% to 10% with two to four genes involved (Ren et al., 2005; Sémon and Duret, 2006; Zhan et al., 2006; Ren et al., 2007). In the fruit fly genome, Spellman and Rubin (2002) found 20% of adjacent genes that were part of large co-expression domains involving ten to 30 genes. In this study, we found that gene islands were significantly enriched in genes showing the same transcription profiles. Indeed, we identified 186 genes (11% of expressed genes in islands *vs.* 4% as expected by random) being part of 80 co-expression clusters of two to 21 genes. It is worth noting that this proportion of co-expressed genes in islands might be even stronger since only one third of the chromosome 3B genes have been studied. In particular, our approach cannot detect tandemly duplicated genes that have been shown as being more prone to co-expression than adjacent non tandemly duplicated genes (Williams and Bowles, 2004; Ren et al., 2005; Zhan et al., 2006). Consistent with this, 89% of the strictly collinear co-expressed genes were shown to be old tandemly duplicated genes. We also found co-expressed gene clusters that were conserved between wheat and rice or *Brachypodium* suggesting that genes were kept close to each other during the evolution to maintain a common regulatory mechanism. Other examples have been already reported of co-expressed and co-functional genes pairs conserved between *Arabidopsis*, rice and poplar (Krom and Ramakrishna, 2008; Liu and Han, 2009). However, beside these collinear genes, we found that more than two third of the co-expression clusters are composed of non collinear genes compared to the rice and/or *Brachypodium* genes. As discussed previously, these clusters likely originate from gene movement mediated by TE or DSB repair and have probably evolved shared expression patterns. This would imply a mechanism by which a

gene, when translocated to a different domain of the genome, may acquire an expression pattern that strongly correlates with the expression patterns of the domain of integration. A similar phenomenon has been previously described by Gierman et al. (2007) who carried out experiments where they inserted a Green Fluorescent Protein (GFP) reporter construct at different chromosomal positions in the human genome. They found that the expression levels of the GFP corresponded to the transcriptional activity of the integration domain, demonstrating that a gene inserted in the vicinity of another one has a significant probability to share the same expression profile than its neighbor. Based on our current data, we cannot estimate the relative contribution of each type of regulation (direct interactions and chromatin effects) on gene co-expression in wheat. Nevertheless, this will be investigated in a near future when we will have access to the chromosome 3B sequence.

Beside these co-expressed genes, we found a significant number of co-functional genes *i.e.* sharing the same GO term(s). Some of them also shared the same expression profiles but the majority was not co-expressed. For the latter, we could hypothesize that we failed in detecting their co-expression as we established the expression profiles on a limited number of cDNA samples. In addition the expression profiles that we established may not reflect the exact expression profiles of the genes located on chromosome 3B since microarray does not allow to discriminate between the three homoeologous copies. However various studies on multiple model organisms demonstrated that clusters were not only found for co-expressed genes, but were also found for genes with products that are involved in the same metabolic pathway or that are associated in protein–protein complexes (Cohen et al., 2000; Lee and Sonnhammer, 2003; Krom and Ramakrishna, 2008; Xu et al., 2008). The mechanisms involved in neighbor gene regulation described above cannot explain the tendency of non co-expressed co-functional genes to occur in adjacent or nearby locations in the genome. To some extent, clusters of co-expressed and/or co-functional genes might be the result of neutral coevolution (Sémon and Duret, 2006). However, assuming that such neutral coevolution would result in a random distribution of co-expression and/or co-function clusters, our results suggest that clusters in wheat do at least partially result from natural selection as they are found more often than expected by chance. Similar conclusions have already been reached in other organisms (for a review, see (Hurst et al., 2004). Therefore, it is very likely that rearrangements in the wheat genome have led to the formation of co-expression and/or co-functional clusters that were maintained in the population due to some selective advantage, as

proposed by Batada et al. (2007), Sémon and Duret (2006) and Madan Babu et al. (2008). The fact that these clusters are more abundant in the distal parts of the chromosome corresponding to the most dynamic regions of the genome (Eichler and Sankoff, 2003; See et al., 2006) suggests a role of this phenomenon in wheat evolution and adaptation.

Positioning of genes within the three-dimensional chromatin structure of the wheat genome might promote long-distance co-regulation of gene clusters

Co-regulation islands, *i.e.* clusters of genes that share the same expression profiles distantly on the same or different chromosomes, have been identified in numerous organisms, including yeast and bacteria (Cohen et al., 2000; Carpentier et al., 2005; Lercher and Hurst, 2006). By establishing a correlation map of the wheat chromosome 3B (Figure 28B), 18 putative intrachromosomal co-regulation islands were identified. These islands involve roughly 70% of the co-expression clusters identified on chromosome 3B, stressing the possible importance of mechanisms underlying long distance regulation of gene expression in wheat. Moreover, since this correlation map was constructed only for chromosome 3B, one cannot exclude the existence of additional interchromosomal co-regulation islands. Hurst et al. (2004) and Madan Babu et al. (2008) described various mechanisms for long-range regulation in three dimensions contrary to the *cis*-regulation of adjacent genes. The formation of DNA loops mediated by specific proteins and the intra- or interchromosomal interactions within the nucleus was shown to have a significant impact on gene *trans*-regulation. Particularly the interactions at the boundaries of chromosomal territories in the centre of the nucleus and near nuclear pores seem to enhance gene expression. Lanctôt et al (2007) also reported an intranuclear polarized organization of the chromosomal segments in mouse. Indeed, the gene-rich regions tend to be oriented towards the center of the nucleus and the gene-poor regions towards the periphery which is less active in terms of transcription. In wheat, the interphase chromosomes adopt a regular Rabl configuration, a highly polarized pattern with the two chromosome arms lying next to each other and the centromeres and telomeres located at opposite poles of the nuclei (Dong and Jiang, 1998; Cowan et al., 2001; Santos et al., 2002). This configuration strikingly departs from the classical cloud-shaped nuclei observed in many organisms during interphase. Beside wheat, the Rabl organization was also observed in large genome species including rye, barley and oats but not in small

genome species such as rice and sorghum (Dong and Jiang, 1998). Consequently, this specific pattern was postulated to be correlated with genome size and chromosome length. In addition, in wheat, because of the increased density of genes towards the telomeres, the Rabl organization has been suggested to play a role in gene expression through the positioning of telomeric gene-rich regions. However, Abranches et al. (1998) clearly demonstrated that active transcription sites were distributed throughout the wheat genome and did not show any preferential localization in the nuclei. In our study, the density of genes involved in co-regulation islands was positively correlated with the distance to the centromere on chromosome 3B. Interestingly, very few of the wheat 3B co-regulation islands involved only collinear genes, meaning that these islands are likely not conserved in rice and *Brachypodium*. The fact that rice (and probably *Brachypodium* as a small genome species) does not display a Rabl configuration suggests that this specific pattern might be a regulatory mechanism in large genome plant species. Based on these findings, we hypothesize that the Rabl organization, if not involved in gene expression *per se*, is likely involved in the co-regulation of distant gene clusters in the wheat genome through the positioning of genes within the nucleus. Chromosome conformation mapping experiments such as *in situ* hybridization or chromosome conformation capture (3C; Shaw, 2010) will be necessary to test this hypothesis.

Here, by constructing the first transcription map of a wheat chromosome, we gained significant insights in the organization, evolution and function of the wheat gene space. Specific features have been identified that strongly suggest that the wheat genome has evolved more rapidly and more drastically than the two model crop species, namely rice and *Brachypodium*. With the development of the other physical maps, similar studies could then be performed at the whole genome scale to confirm the features observed on chromosome 3B. Finally, with the sequencing of chromosome 3B underway, we should be able to refine our knowledge and answer the different questions raised by this study.

METHODS

The wheat NimbleGen 40K unigene microarray

The wheat NimbleGen 40K unigene microarray was designed using the *Triticum aestivum* NCBI unigene set build #55 counting 40349 unigenes resulting from the assembly of 960174 sequences (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=4565).

Unigene sequences were masked based on a k-mer frequency analysis. A 17-mer-based mathematically defined repeat (MDR) index was build with Tallymer (Kurtz et al., 2008) by using 2 Gb of Illumina reads produced from sorted 3B chromosomal DNA (accession number: ERA000182). K-mer frequency of all unigene sequences was thus computed using this MDR index. Seventeen-mers repeated five times or more times in the index were masked to exclude repeated motifs from the probe design. The masked sequences were submitted to NimbleGen (Roche NimbleGen, Inc.) for probe design using proprietary algorithms. In total 39179 unigenes were represented by at least one 60-mer probe: 39019 unigenes with three probes, 78 with two probes and 82 with one probe. These probes were synthesized onto 12×135K array.

Construction of the second version of the wheat chromosome 3B physical map and production of MTP BAC pools

A new chromosome 3B-specific BAC library containing 82,176 clones was constructed as described by Safar et al. (2004). The 82,176 BAC clones together with the 7440 MTP BAC clones originating from the first version of the physical map were fingerprinted using a slightly modified High Information Content Fingerprinting (HICF) SNaPshot protocol that uses a combination of five Type II restriction enzymes and capillary electrophoresis on automated sequencers (Paux et al., 2008). A total of 78,840 new high quality fingerprints were obtained and analyzed with Fingerprinted Contigs Program (FPC) (Soderlund et al., 1997; Soderlund et al., 2000) for being included into the first version of the chromosome 3B physical map containing 1036 contigs. Briefly, the initial build of chromosome 3B was performed by incremental contig building with a cutoff of 1e-75 and a tolerance of 4. These were subsequently run through keyset-to-FPC, single-to-end and end-to-end merging (Match:

1; FromEnd: 55) at six successively higher cutoffs terminating at 1e-45. The DQer function was used after each merge to break up all contigs that contained more than 10% of Questionable (Q) clones (Step: 3). Sixty-four three-dimension (plate, row, column) MTP BAC pools were produced following the procedure described by Paux et al. (2008).

Microarray hybridizations

Total RNAs of hexaploid wheat *cv* Chinese Spring were extracted in duplicate from five organs (root, leaf, stem, spike and grain) at three developmental stages each (beginning, middle and end of the development) and were reverse transcribed as detailed by Choulet et al. (2010).

The 64 MTP pools were sonicated to obtain an average fragment size between 500 and 2000 bp. The cDNA and the MTP pools labellings were carried out using the NimbleGen Dual-Color DNA Labeling Kit (Roche NimbleGen, Inc.) according to the manufacturer's procedure for gene expression analysis. Dual color hybridizations were performed on each plex of the arrays. The labelled cDNAs and MTP pools were hybridized independently. The pairs of samples were randomly chosen and allocated to a plex. A dye swap was performed for all the duplicates cDNAs making four repetitions for each of the 15 samples. Hybridizations and washing of the arrays were performed according to the manufacturer's procedure for gene expression analysis (Roche NimbleGen, Inc.). The arrays were scanned using the InnoScan 900AL scanner (Innopsys, Carbonne, France). Data was extracted from scanned images using NimbleScan 2.5 software (Roche NimbleGen, Inc.) which allows for automated grid alignment, extraction and generation of data files.

Normalization and data deconvolution of the MTP pool data

The normalization and the data deconvolution were performed using automated scripts developed with the R software (www.r-project.org). First intensity values were checked for each gene and each MTP pool. Probes showing aberrant signal intensity were deleted. Data from each MTP pool were then made comparable with each other by subtracting the median to each intensity value and then by dividing by the standard deviation. Three files were

generated corresponding to the normalized intensities for the plate, row and column pools. The three pool types were treated separately. Two complementary methods with three stringency thresholds each were then used to detect the positive signals.

The first method, called The "Mean + X × Standard Deviation" method, required the calculation of the median for the intensities of probes corresponding to the same gene and to the same pool. Then the same method than the "automated scoring" method described by Rustenholz et al. (2010) was applied to the data with minor modifications. Indeed the coefficients multiplying the standard deviation were modulated and the plate and column pools shared the same coefficients (Coefficients for high thresholds: plate and column=2.8 and row=2.5; Coefficients for medium thresholds: plate and column=2.4 and row=2.3; Coefficients for low thresholds: plate and column=2.2 and row=2.1).

The second method, called the "t-test" method, used Student's t-Tests for each gene to compare intensities of one pool to the intensities of the others. Variance was considered to be equal. The three stringency levels corresponded to the three *P*-value thresholds (High, *P*-value threshold=0.01; Medium, *P*-value threshold=0.025 and Low, *P*-value threshold=0.05).

Data deconvolution was carried out independently for both methods and stringency levels as described by Rustenholz et al. (2010). All BAC addresses found with the high thresholds of both methods were retained. Then for the medium thresholds, only the unigenes located on the overlap of two BACs and that were not previously identified by the high thresholds were added. Finally for the low thresholds, only the unigenes located on the overlap of two BACs and that were not previously identified by the high and medium thresholds were added.

Sequence analyses

Since EST assembly used for the unigene design is not available at the NCBI web site, we rebuilt the EST contigs to check if some could exhibit highly similar regions. FASTA sequences of ESTs belonging to the same cluster have been pooled together and assembled using Phrap (http://www.phrap.org/) to rebuild of the EST contigs. Two rounds of assembly were performed: first, using default Phrap parameters; second, with relaxed stringency in order to assemble EST clusters for which 2 or more contigs were obtained after the first

round. Then, similarity search between EST contigs was conducted using BLASTN (Altschul et al., 1997). Alignments were parsed to keep only the best hits with at least 90% of sequence identity on at least 20 bp. Pairwise comparisons of EST contigs were performed and unigenes showing at least 98% of sequence identity, mapped to the same BAC clone and potentially matching the same gene were removed from further analyses. BLASTN (Altschul et al., 1997) analyses of the unigene sequences provided by the NCBI and of the sequences of the probes designed within the unigene sequences against the sequenced contigs of chromosome 3B (Choulet et al., 2010) were performed. The results were parsed to keep the hits for which the query sequence matched on at least 20% of its length for the unigene sequence and 40% of its length for the probes sequences. The Artemis viewer (Rutherford et al., 2000) was used to found the annotation (gene, non annotated region or repeat) at the hit position. BLASTX (Altschul et al., 1997) analyses of the unigene sequences against the databases of all the rice peptides (http://rice.plantbiology.msu.edu) and of all the *Brachypodium* peptides (http://www.brachypodium.org) were performed to identify to orthologs in the wheat genome. The results were parsed to keep the best hits with at least 50% of sequence similarity on at least 33 amino acids. Every unigene with a hit meeting these criteria was considered as orthologous the rice or *Brachypodium* gene identified. Two genes on the same wheat contigs were considered as collinear if their orthologs in the rice or *Brachypodium* genome were separated by less than ten genes. The GO annotations from rice and from *Brachypodium* (ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/ontology/go/) were used to assess the GO of the wheat genes orthologous to rice and/or *Brachypodium* genes. A GO-Slim annotation (goslim_plant.obo selected) was performed using Blast2GO (http://www.blast2go.org/).

A BLASTN (Altschul et al., 1997) analysis of the barley unigene set used by Rustenholz et al. (2010) against the NCBI wheat unigene set build #55 was performed. The results were parsed to keep only the best hits with at least 80% of sequence identity on at least 100 bp. Every unigene with a hit meeting these criteria was considered as orthologous between wheat and barley.

Transcriptomic data analyses and validation

The normalization of the transcriptomic data was performed using automated scripts developed with the R software (www.r-project.org). First each image was divided into 30 identical zones. The background intensity was estimated using the median of the intensities of the empty spots and of the spots made of random sequences for each zone and subtracted to each spot intensity within the zone. Then two successive lowess corrections were undertaken to correct the dye and the duplicate biases. Afterwards the corrected intensity values were checked for each gene and the aberrant values were deleted. Student's t-Tests were performed to compare the corrected intensity values for each gene with the corrected background intensity values. The *P*-value threshold was set at 0.05. The median of the corrected intensity values was calculated for each significantly expressed gene and was considered as the expression value of the gene for the cDNA sample. To harmonize the expression values between the 15 cDNA samples, the expression values of each significantly expressed gene were subtracted by the median of the expression values of the significantly expressed genes for the cDNA sample and then divided by the standard deviation. All the expression values were finally made positive and the expression value of a gene that was not significant expressed in a cDNA sample was set to 0.

The validation of the expression profiles established through microarray hybridization was carried out on eight genes. The QRT-PCR experiments were performed on the 30 cDNA samples (15 samples extracted in duplicate) in duplicate on the LightCycler 480 (Roche Diagnostics) using the LightCycler 480 SYBR Green I Master mix (Roche Diagnostics) following the manufacturer's procedure. The data were analyzed using the LightCycler480 Software release 1.5.0 (Roche Diagnostics) with the Absolute Quantification / Fit Points method. The median of the four values obtained for each sample was calculated to be compared with the expression data from microarray hybridization.

Hierarchical clustering was performed using the Hierarchical Clustering Explorer 3.0 software (http://www.cs.umd.edu/hcil/hce/hce3.html) with the complete linkage method and the Pearson correlation coefficient. The minimal similarity to establish the clusters was set to 0.641 which is a Pearson correlation significant at the *P*-value threshold of 0.01.

We used the rice transcriptomic data developed by Wang et al. (2010) as they sampled the same organs than the one we selected, on the entire life cycle of the plant. Medians were calculated for the expression data corresponding to root, leaf, stem, spike and grain parts.

Statistical analyses

The statistical analyses were performed using automated scripts developed with the R software (www.r-project.org).

The script developed to validate the wheat gene space organization in gene islands performed 10,000 random samplings with replacements of 2924 BACs out of all the MTP BACs of wheat chromosome 3B. It then calculated the percentage of identical and overlapping BACs retrieved. The other scripts used to validate the co-expression level, the co-function level and the co-function level of the co-expressed genes performed 10,000 random samplings without replacements of the 2924 chromosome 3B BAC addresses where a gene was located. They then calculated either the percentage of co-expressed genes retrieved at the gene island scale or the percentage of co-functional genes retrieved at the gene island scale or the percentage of co-expressed and co-functional genes retrieved at the gene island scale.

Chi² tests were performed to check the uniformity of a gene density distribution along wheat chromosome 3B. First the average density was calculated by dividing the sum of genes (isolated or co-expressed genes) by the total length of contigs assigned to the deletion bins. Then the number of genes per deletion bin under the uniform law was calculated by multiplying the average density with the length of contigs assigned per deletion bin. The numbers of genes observed for all the deletion bins and the numbers of genes for all deletion bins under the uniform law were compared through a Chi² test. The *P*-value threshold was set at 0.05.

Classical Pearson's correlation coefficient tests were performed to check the correlation between various variables. The *P*-value threshold was set at 0.05.

Correlation matrixes were established by calculating Pearson's correlation coefficient of the expression profiles between all the pairs of genes on wheat chromosome 3B or on a specific region of rice chromosome 10 (LOC_Os10g21194 to LOC_Os10g38276). The wheat genes

were ordered thanks to their assignation to a deletion bin. The wheat contigs within the deletion bins were randomly ordered. The correlation maps were established based on the correlation matrixes using the corrplot package (http://cran.r-project.org/).

ACKNOLEDGMENTS

The authors would like to thank Isabelle Bertin and Nelly Cubizolles for their technical assistance with the QRT-PCR experiment and Pierre Sourdille for helpful comments and discussions. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement n°FP7-212019 and from the Institut National de la Recherche Agronomique (AIP "ChromBlé"). Camille Rustenholz was financially supported by Région Auvergne.

Conclusion Article n°3

La cartographie d'environ 3000 gènes sur le chromosome 3B nous a permis d'étudier l'organisation de l'espace génique sur un chromosome de blé avec un nombre de gènes et une résolution jamais atteinte jusqu'à présent. Les résultats obtenus confirment que les gènes sont organisés selon un gradient de densité du centromère vers les télomères le long du chromosome 3B. De plus, 70% des gènes ont été identifiés comme faisant partie d'îlots de gènes et la distribution de ces gènes expliquerait le gradient observé.

Les analyses évolutives des gènes portés par le chromosome 3B du blé par rapport aux gènes de riz et de *Brachypodium* ont montré que la distribution des gènes en îlots coïncidait avec celle des gènes dits non synténiques ou réarrangés. Ainsi les îlots seraient plutôt générés par le déplacement ou la copie de gènes à proximité d'autres gènes très probablement via des ET ou des évènements de cassures double-brin au moment de la recombinaison.

Une analyse de l'expression des gènes positionnés sur le chromosome 3B a permis de mettre en évidence des clusters de gènes coexprimés partageant également la même fonction pour la majorité d'entre eux. De nombreux autres îlots de gènes partageant la même fonction ont été identifiés le long du chromosome 3B sans être associés à la coexpression. Ainsi nous avons démontré que dans le génome du blé, comme dans d'autres espèces, les gènes avaient des raisons fonctionnelles d'être gardés proches les uns des autres.

Finalement, nous avons émis l'hypothèse que les îlots de gènes observés le long des chromosomes du blé étaient la résultante de mécanismes structuraux et fonctionnels. En effet ceux-ci seraient formés suite des déplacements de gènes via des ET puis seraient conservés et sélectionnés en partie pour des raisons fonctionnelles de coexpression ou de cofonction. De plus, des mécanismes de régulation entre groupes de gènes distants portés par le chromosome 3B ont été suspectés.

CONCLUSIONS

ET

PERSPECTIVES

Les travaux présentés dans cette thèse ont permis d'affiner notre connaissance de la structure du génome du blé et plus particulièrement de l'organisation de l'espace génique. Dans un premier temps, l'obtention de résultats fiables a permis de valider notre stratégie utilisant des outils de transcriptomique, tels que les puces à ADN, pour cartographier les gènes sur les BAC du chromosome 3B. En effet, après avoir hybridé des ADNc sur des BAC, la technique consistant à hybrider des pools de BAC sur des puces portant des unigènes s'est avérée une approche fiable, efficace et à relativement haut-débit pour positionner des gènes sur des BAC. Nous avons ainsi cartographié presque 3000 gènes le long d'un chromosome de blé à l'échelle du BAC, soit un nombre jamais atteint jusqu'à présent.

La connaissance précise de la position des BAC grâce à l'établissement d'une nouvelle version de la carte physique couvrant 97% du chromosome 3B a, par la suite, permis de mener une analyse approfondie de la distribution de ces gènes le long du chromosome. Ainsi, les gènes sont répartis tout le long du chromosome 3B sans aménager de régions de plusieurs mégabases totalement dépourvues de gènes. Cependant ces gènes ne sont pas répartis de façon homogène le long du chromosome puisqu'un gradient positif de la densité de gènes du centromère vers les télomères a été identifié. Nous avons estimé que les régions télomériques portaient environ deux fois plus de gènes que les régions centromériques. Cette analyse à l'échelle d'un chromosome entier de blé a été un argument en faveur d'un séquençage complet des chromosomes puisque le séquençage ciblé sur les BAC « riches en gènes » identifiés par criblage de banques BAC avec des EST n'auraient jamais permis de capturer la totalité des gènes.

A une échelle plus fine, l'analyse du nombre de gènes présents sur le même BAC ou sur les BAC chevauchants a permis d'estimer que 70% des gènes étaient organisés en îlots sur le chromosome 3B. Ainsi, cette analyse de l'espace génique du blé à l'échelle du BAC a permis de confirmer que les gènes étaient majoritairement organisés en îlots de deux à quatre gènes et que d'autres étaient isolés au sein d'océans d'ET insérés les uns dans les autres. De plus, la densité de gènes en îlots le long du chromosome 3B a permis d'expliquer le gradient positif de gènes puisque les gènes en îlots ont une densité significativement plus importante dans les régions distales que dans les régions proximales. Nous sommes donc parvenus à confirmer pour la première fois et grâce à un nombre important de gènes que la répartition hétérogène des gènes à l'échelle du chromosome 3B entier était le reflet d'une répartition hétérogène des gènes à l'échelle de la séquence.

Cependant la plus grande originalité de mon travail de thèse a consisté à initier des réponses aux questions concernant la formation et la fonctionnalité des îlots de gènes et de leur gradient de distribution. D'un point de vue évolutif, nous avons mis en évidence que la répartition des gènes conservés entre le blé et le riz était également hétérogène et suivait un gradient inversé par rapport à la densité de gènes le long du chromosome. Par contre la densité de gènes dits réarrangés était positivement corrélée à la densité de gènes et à la densité de gènes en îlots. Ainsi, l'hétérogénéité de la répartition des gènes réarrangés expliquerait également le gradient de la densité de gènes le long du chromosome 3B du fait de l'enrichissement en gènes réarrangés dans les îlots de gènes surtout dans les régions télomériques.

Du point de vue fonctionnel, les îlots de gènes sont significativement enrichis en gènes proches partageant des profils d'expression similaires, des fonctionnalités communes ou les deux. De plus, la majorité des clusters de coexpression était constituée de gènes réarrangés et un enrichissement en gènes coexprimés a été observé au niveau des télomères. Ces observations ont permis d'émettre une hypothèse quant à la formation et au maintien des îlots de gènes essentiellement au niveau des télomères. Les îlots de gènes se seraient formés par des réarrangements de gènes initiés par l'insertion d'ET et des cassures double-brin dans les régions distales. Comme l'insertion d'un gène à proximité d'un autre favoriserait la coexpression du fait d'un partage de la structure chromatinienne entre ces deux gènes, des clusters de gènes coexprimés nouvellement formés ne seraient maintenus que s'ils apportaient un avantage adaptatif à l'individu. D'après nos observations, les clusters de gènes coexprimés sont également très largement cofonctionnels, ce qui suggère que les gènes, réarrangés à proximité de gènes partageant la même fonctionnalité, ont formé des îlots qui seront conservés préférentiellement du fait de leur coexpression.

Enfin, grâce à l'analyse de l'expression des gènes positionnés sur le chromosome 3B, des régions partageant probablement des mécanismes de régulation communs à distance ou îlots de corégulation ont été mis en évidence pour la première fois dans le génome du blé. Ces interactions suggèrent ainsi un ordre supérieur en trois dimensions de la régulation des gènes.

Cette thèse a ainsi permis de préciser l'organisation de l'espace génique mais aussi d'ouvrir des perspectives vers la compréhension de la formation, du maintien et de la fonctionnalité de l'hétérogénéité de la distribution des gènes de blé le long des chromosomes à différentes échelles.

1. L'UTILISATION DE LA SEQUENCE DU CHROMOSOME 3B ENTIER POUR ANALYSER ET COMPRENDRE L'ORGANISATION DE L'ESPACE GENIQUE

Avec l'utilisation de la carte physique couvrant 97% du chromosome 3B, 3000 gènes ont été cartographiés le long du chromosome ce qui n'avait jamais été réalisé jusqu'à présent. Cet outil nous a ainsi permis d'étudier l'organisation de l'espace génique avec une grande précision. Cependant plusieurs limites techniques de cette étude n'ont pas rendu possible l'analyse tout à fait exhaustive de l'organisation de l'espace génique du chromosome 3B. Tout d'abord, Choulet et al. (2010) ont estimé que le chromosome 3B comptait environ 8400 gènes alors que notre stratégie basée sur l'hybridation n'a permis d'en cartographier que 3000. De plus, la résolution à l'échelle du BAC ne nous a pas permis d'affirmer avec certitude que les gènes positionnés sur le même BAC formaient un îlot et de savoir si les gènes présentaient des copies dupliquées en tandem sur le même BAC.

Ainsi comme suggéré par Choulet et al. (2010), seul le séquençage du chromosome entier permettra d'aboutir à la résolution nécessaire pour, d'une part, avoir accès à l'ensemble des gènes portés par le chromosome 3B, gènes dupliqués en tandem inclus, et pour, d'autre part, estimer précisément les distances intergéniques et donc étudier les îlots de gènes réels. En conséquence, une séquence de bonne qualité, avec une bonne couverture, assemblée et annotée avec précision, permettrait d'appréhender l'organisation de l'espace génique du chromosome 3B de la façon la plus exhaustive possible.

1.1. L'analyse exhaustive de l'organisation de l'espace génique grâce à la séquence complète du chromosome 3B

Dans le cadre du projet 3BSEQ (ANR-projet phare 2009-2012), le séquençage du chromosome 3B du blé tendre est actuellement en cours au Génoscope. Les contigs de BAC de la dernière version de la carte physique sont séquencés avec une couverture moyenne de 40X. Les séquences assemblées des contigs seront ensuite annotées de façon automatisée dans le courant de l'année prochaine avec le pipeline d'annotation Triannot développé dans l'équipe. Cette annotation ciblera aussi bien les gènes que les séquences répétées. En

parallèle, les séquences seront ordonnées le long du chromosome par la combinaison de différentes approches de cartographie (population recombinantes, lignées aneuploïdes, panel d'hybrides de radiation…). Ainsi une analyse exhaustive et précise de l'organisation de l'espace génique du chromosome 3B sera possible dès l'année prochaine.

1.1.1. La caractérisation structurale exhaustive de l'organisation de l'espace génique

La séquence complète du chromosome 3B fournira la vision la plus précise possible de la structure de ce chromosome aussi bien du point de vue des gènes que du point de vue des séquences répétées. Les observations réalisées sur les contigs séquencés (Choulet et al., 2010) pourront ainsi être vérifiées et extrapolées à l'ensemble du chromosome.

Dans un premier temps, la densité de gènes moyenne du chromosome 3B pourra être établie et comparée avec des densités de gènes locales. Ainsi, d'éventuelles régions riches et pauvres en gènes pourront être identifiées et l'hypothèse de la distribution des gènes selon un gradient positif de la densité de gènes du centromère vers les télomères pourra être définitivement vérifiée.

De plus, la séquence du chromosome 3B permettra de réaliser une étude approfondie des gènes dupliqués en tandem le long du chromosome 3B. Ainsi, l'hypothèse émise par Akhunov et al. (2003) et Choulet et al. (2010) concernant le gradient positif de gènes dupliqués en tandem du centromère vers les télomères pourra être testée. En effet, si cette hypothèse est correcte, les gènes dupliqués en tandem renforceraient encore le gradient de gènes au niveau de régions télomériques comme estimé dans l'article n°3.

En outre, les analyses réalisées par Choulet et al. (2010) sur les distances intergéniques observées dans les contigs séquencés, pourront être étendues à l'ensemble du chromosome 3B grâce à la séquence. La distribution des distances intergéniques de l'ensemble du chromosome permettra d'affiner la définition des îlots de gènes chez le blé et donc de déterminer avec précision la proportion de gènes en îlots par rapport à celle des gènes isolés. Par la suite, l'analyse de la distribution des gènes en îlots le long du chromosome 3B permettra de vérifier leur implication dans le gradient de gènes le long du chromosome si celui-ci est validé.

Au niveau des îlots de gènes, des analyses approfondies pourront être menées pour tenter d'identifier des séquences régulatrices communes à l'ensemble des gènes de l'îlot. De plus, l'orientation des gènes, dont l'impact a pu être démontré sur la régulation des gènes proches (Zhan et al., 2006; Chen et al., 2010a), pourra également être analysée avec précision.

1.1.2. L'analyse du degré de conservation des gènes portés par le chromosome 3B

Outre les analyses structurales, les analyses évolutives des gènes portés par le chromosome 3B initiées lors de ma thèse pourront également être considérablement approfondies. L'utilisation des séquences complètes et précises des gènes de blé lors de la recherche de gènes orthologues chez différentes espèces facilitera l'identification des gènes synténiques et des gènes réarrangés. Ainsi les analyses des relations entre l'organisation des gènes et leur conservation pourront encore être approfondies aussi bien à l'échelle du chromosome entier que des îlots de gènes. Elles permettront de valider l'hypothèse énoncée par Choulet et al. (2010) et dans l'article n°3, concernant l'impact majeur des gènes réarrangés dans la formation des îlots de gènes principalement au niveau des régions télomériques du chromosome 3B.

De plus, grâce à la séquence du chromosome 3B, il sera également possible de vérifier si les gènes non synténiques identifiés ont été déplacés par des mécanismes proposés par Wicker et al. (2010). Ces derniers suggèrent que la capture de fragments ou de gènes entiers dans les ET au moment de leur transposition serait un mécanisme mineur dans la translocation ou la duplication de gènes entiers. Ils émettent l'hypothèse que les cassures double-brin générées suite à l'insertion d'ET sont parfois réparées en utilisant des fragments de séquence provenant d'autres chromosomes suffisamment grands pour contenir un ou plusieurs gènes entiers. Wicker et al. (2010) ont établi cette hypothèse en se basant sur des séquences de quelques nucléotides retrouvées en amont de l'ET inséré et en aval du fragment dupliqué. De telles séquences pourront ainsi être recherchées aux alentours des gènes réarrangés dans la séquence du chromosome 3B. De plus, une attention particulière pourra être portée à la séquence portant une vingtaine de gènes réarrangés par rapport au génome du riz décrite dans l'article n°3. L'analyse de cette séquence permettra peut-être de savoir si les mécanismes énoncés par

Wicker et al. (2010) peuvent s'étendre à plusieurs dizaines de gènes ou si un autre mécanisme a généré ce grand déplacement de gènes en groupe.

1.1.3. L'annotation fonctionnelle des gènes portés par le chromosome 3B

Enfin, en plus de l'analyse structurale exhaustive et de l'analyse des relations entre l'organisation de l'espace génique et la conservation des gènes, la séquence des gènes du chromosome 3B permettra également de réaliser leur annotation fonctionnelle précise. Ainsi, il sera possible de mener une étude beaucoup plus approfondie des relations entre l'organisation de l'espace génique et la fonction des gènes à différentes échelles. Des régions enrichies en gènes de fonction particulière pourront peut-être être mises en évidence sur le chromosome 3B comme des clusters de gènes de ménage (Lercher et al., 2002; Liu and Han, 2009). De plus, à l'échelle des gènes, l'enrichissement des îlots de gènes en gènes cofonctionnels pourra être vérifié.

1.2. La quantification précise de l'expression des gènes par RNAseq

Dans le cadre du projet 3BSEQ, l'utilisation de la technologie RNAseq sur les 15 échantillons d'ARN décrits dans les articles n°1 et 3 est également planifiée. Le RNAseq utilise les nouvelles technologiques de séquençage à haut débit pour générer des lectures courtes (25-150 nt) mais avec une couverture importante de banques d'ARNm ou d'ADNc. Ces millions de séquences courtes doivent ensuite être alignées contre un génome de référence pour faciliter la quantification de l'expression des gènes réalisée par simple comptage du nombre de lectures par gène. Ainsi, cette technique de quantification de l'expression a été reconnue comme étant extrêmement fiable par comparaison avec les techniques de quantification plus traditionnellement utilisée à l'heure actuelle comme les puces d'expression et la Q-PCR (Cloonan and Grimmond, 2008).

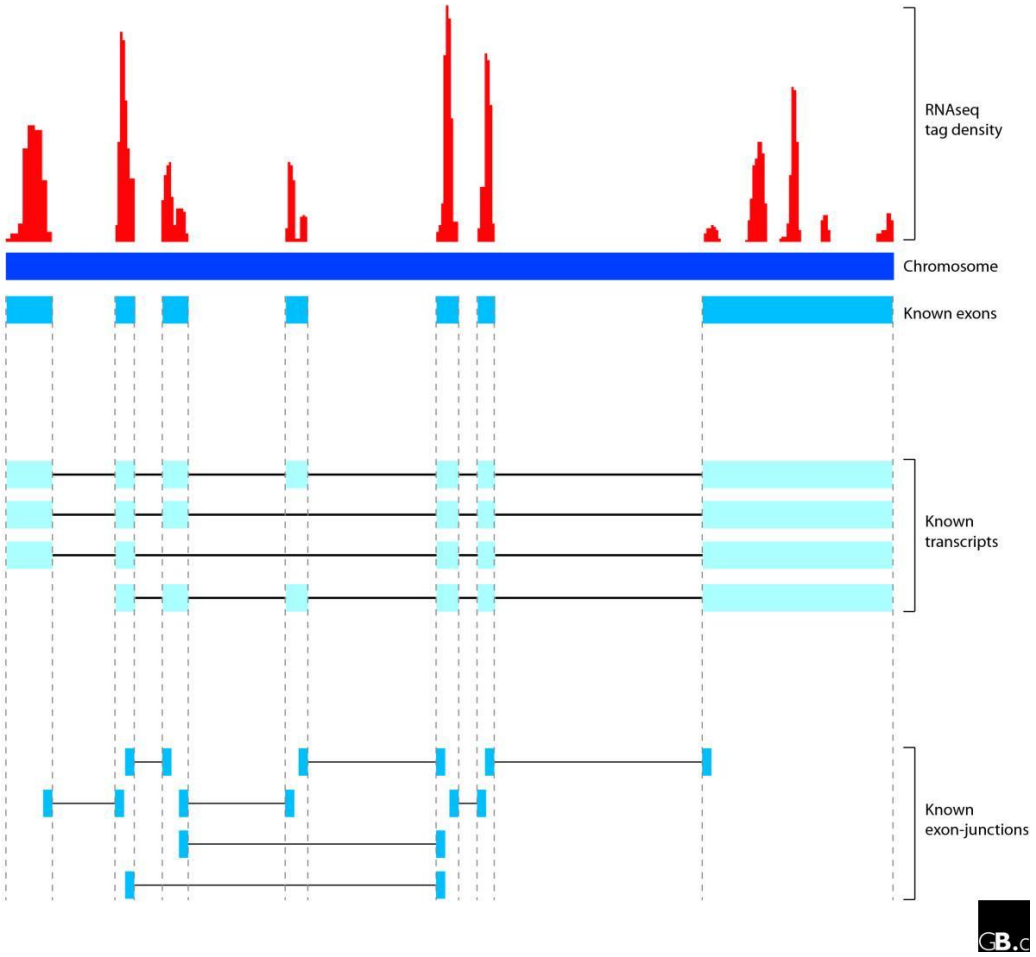


Figure 29 : Le RNAseq comme outil d'aide à l'annotation. Sur ce schéma, les séquences issues du RNAseq (en rouge) sont alignées sur la séquence du génome ce qui permet une visualisation quantitative de la densité de séquences RNAseq au niveau de ce locus. Cet alignement contre le génome permet d'identifier les bornes exon-intron et de détecter de l'épissage alternatif. (d'après Cloonan et Grimmond, 2008)

1.2.1. Le RNAseq : un outil d'aide à l'annotation

Dans le cadre du projet 3BSEQ, le RNAseq sera utilisé pour apporter des preuves biologiques comme soutien à l'annotation des gènes du chromosome 3B. En effet, comme présenté par Cloonan et Grimmond (2008) (Figure 29), cet outil permettra de préciser les annotations des bornes exon-intron et peut-être aussi d'observer de l'épissage alternatif. Le RNAseq apportera donc des preuves biologiques à l'annotation réalisée.

De plus, à la différence des puces d'expression, le RNAseq est une technique de transcriptomique sans a priori. Ainsi, l'utilisation de cette technologie sera probablement l'occasion de découvrir de nouvelles unités transcriptionnelles comme de nouveaux gènes ou pseudogènes qui n'auraient pas été annotées faute de preuves biologiques ou de structure génique caractéristique. En effet, de telles régions ont déjà été repérées lors de l'analyse développée dans l'article n°3 et leur identification pourrait être facilitée par l'analyse des séquences produites par RNAseq. De plus, des régions générant des transcrits non codants impliqués dans la régulation pourraient être découvertes (Lister et al., 2008). Des transcrits issus d'ET ou chimériques du type ET-gène (Lister et al., 2008) ainsi que des transcrits naturels en antisens (cis-NAT; Conley et al., 2008; Jin et al., 2008) pourraient également être identifiés. Enfin, si le RNAseq est réalisé à partir de banques d'ARN, des séquences de petits ARN ou ARN sans queue polyA pourraient être identifiées et des zones de production localisées sur le chromosome 3B (Lister et al., 2008). Ainsi cette expérience de RNAseq permettra d'affiner notre connaissance du niveau d'activité transcriptionnelle et de certains mécanismes de régulation du génome du blé de façon exhaustive et à l'échelle d'un chromosome entier.

1.2.2. L'établissement de la carte transcriptionnelle fine du chromosome 3B entier

Une limite à l'utilisation de puces portant des unigènes notamment pour les analyses d'expression est qu'elles ne permettent pas de distinguer l'expression des gènes homéologues. Le génome du blé tendre étant un hexaploïde récent, la majorité des gènes sont susceptibles d'être présents en trois copies homéologues potentiellement fonctionnelles. Comme leurs séquences sont restées très similaires du fait de la polyploïdisation récente, les ADNc

correspondant aux trois copies homéologues se sont probablement hybridés sur le même unigène. Ainsi, les expressions des unigènes obtenues reflètent vraisemblablement les expressions des trois copies homéologues simultanément. De ce fait, mêmes si les résultats sont globalement fiables (cf. article n°3), la carte transcriptionnelle établie par hybridation correspond à l'analyse des relations entre la position des gènes portés par le chromosome 3B et l'expression cumulée des gènes homéologues du groupe de chromosomes 3. Ainsi, certains clusters de gènes coexprimés pourraient être des faux-positifs et d'autres pourraient avoir échappé à notre analyse.

De plus, les EST portées par la puce NimbleGen de blé et sélectionnées par le NCBI pour représenter les unigènes du blé sont issues de multiples banques de transcrits probablement générées à partir de variétés différentes. Or l'ensemble de mon travail est centré autour de la variété de référence Chinese Spring. Bien que le polymorphisme de séquence génique soit restreint entre variétés de blé tendre (Ravel et al., 2006), des divergences de séquences ont pu altérer la quantification de l'expression de certains gènes.

La mise en œuvre de la technique RNAseq pourrait permettre de quantifier l'expression des gènes homéologues indépendamment à partir de la variété Chinese Spring. En effet, en utilisant des lectures de 75 nt pour augmenter la spécificité de chaque séquence, il serait possible de reconstituer les trois groupes d'homéologie (à condition que des polymorphismes puissent être détectés dans ces lectures courtes), d'obtenir la séquence de chacune des copies homéologues et de quantifier l'expression de chacune des copies à partir de leur nombre respectif de lectures. Cependant la limite de cette technologie de transcriptomique sera vraisemblablement atteinte pour des gènes peu exprimés pour lesquels le nombre de lectures ne permettra peut être pas de distinguer les copies homéologues.

En associant le séquençage et l'annotation des gènes du chromosome 3B avec la quantification précise de leur expression par RNAseq, une carte transcriptionnelle fine du chromosome 3B pourra être établie au plus proche de la réalité. Elle permettra d'analyser les relations entre l'organisation de l'espace génique, l'évolution, la fonction et l'expression des gènes portés par le chromosome 3B. Le niveau d'expression des gènes pourra être analysé à l'échelle du chromosome entier et d'éventuelles régions fortement ou faiblement exprimées pourront être identifiées comme dans le génome d'*Arabidopsis* où les régions centromériques correspondent aux régions les moins actives au niveau transcriptionnel (Schmid et al., 2005).

Une attention particulière pourra être portée à l'étude de l'implication des ET sur la régulation de l'expression des gènes dans leur voisinage proche. En effet, de nombreuses études ont démontré le rôle des ET aussi bien dans l'initiation de la transcription de certains gènes (Casacuberta and Santiago, 2003; Kashkush and Khasdan, 2007) que dans leur répression du fait du silencing épigénétique exercé par l'hôte pour contenir la transposition des ET (Hollister and Gaut, 2009). Ainsi, l'impact de la nature des ET à proximité des gènes et des distances les séparant sur le niveau d'expression de gènes proches pourra être étudié finement.

En outre, comme cette carte transcriptionnelle fine sera établie avec tous les gènes portés par le chromosome 3B, l'analyse précise des îlots de coexpression sera facilitée. Une étude approfondie de ces derniers pourra être réalisée pour démontrer l'impact de l'orientation des gènes, des distances intergéniques et des promoteurs sur la coexpression et la régulation de l'expression des gènes adjacents, comme de nombreuses études l'ont déjà fait sur des espèces modèles (Zhan et al., 2006; Krom and Ramakrishna, 2008; Liu and Han, 2009; Chen et al., 2010a).

Enfin, l'établissement de la carte transcriptionnelle du chromosome 3B entier avec des profils d'expression précis permettra de valider la présence d'îlots de corégulation portés par le même chromosome. Ceux-ci pourront alors être étudiés plus spécifiquement afin de comprendre les mécanismes de régulation à longue distance impliqués (cf. §4).

En conclusion, l'établissement de la carte transcriptionnelle fine du chromosome 3B entier avec sa précision et son exhaustivité, devrait répondre au grand nombre d'hypothèses que mon travail de thèse a soulevées. Cependant les analyses seront toujours limitées au chromosome 3B et les aspects de régulations et d'interactions interchromosomiques ne pourront être abordés qu'après le développement d'outils similaires sur les autres chromosomes du blé.

1.2.3. L'impact de la polyploïdie sur l'expression des gènes homéologues

Comme précisé précédemment, la résolution du RNAseq permettra de quantifier l'expression des gènes homéologues individuellement. Outre l'établissement de la carte transcriptionnelle

fine du chromosome 3B, cette technologie permettra également d'analyser l'impact de la polyploïdie sur la régulation de l'expression des gènes homéologues. Par exemple, la répression de l'expression d'une ou de deux copies au profit d'une autre pourrait être observée. De plus, la genèse de transcrits dégénérés pourrait traduire la pseudogénisation d'une copie homéologue. En outre, certaines copies pourraient avoir acquis une spécificité d'expression dans un organe ou à un stade de développement particulier par rapport aux autres copies, ce qui traduirait une subfonctionnalisation. Enfin, une modification conjointe de séquence et de profil d'expression pourrait également impliquer la néofonctionnalisation d'une copie homéologue par rapport aux deux autres. Ainsi, la technologie du RNAseq permettrait de mener l'analyse la plus exhaustive à l'heure actuelle du devenir des gènes homéologues du point de vue de leur expression.

2. LES ANALYSES APPROFONDIES DES GENES REARRANGES ET DE LA FORMATION DES ILOTS DE GENES

L'article n°3 et celui de Choulet et al. (2010) soulignent l'importance des gènes dit réarrangés ou non synténiques, c'est-à-dire ayant subi des réarrangements tels que des translocations ou des duplications et les gènes spécifiques au blé qui n'ont pas pu être identifiés dans les génomes du riz et de *Brachypodium*. Ces gènes seraient responsables du gradient de la densité de gènes le long du chromosome 3B et ils constitueraient la majeure partie des îlots de gènes et des clusters de gènes coexprimés. Grâce à la carte transcriptionnelle fine du chromosome 3B entier, ces hypothèses pourront être vérifiées.

Cependant, Choulet et al. (2010) et Wicker et al. (2010) proposent l'hypothèse que les gènes non synténiques seraient majoritairement issus de duplications interchromosomiques. Ainsi les gènes à l'origine de ces gènes non synténiques seraient toujours en position orthologue par rapport au génome du riz ou de *Brachypodium*. Seules les copies auraient été déplacées ou créées sur d'autres chromosomes et préférentiellement au niveau des parties distales des chromosomes. Il serait ainsi intéressant de vérifier que les gènes à l'origine des gènes non synténiques ont été majoritairement conservés en position orthologue. Dans un premier temps, une expérience de Q-PCR pourrait être mise en œuvre sur le Biomark (Fluidigm Corporation, San Francisco, USA) disponible au laboratoire pour évaluer le nombre de copies de ces gènes

dans le génome de façon très résolutive. En effet, la sensibilité de cet appareil permet de distinguer une différence d'une copie entre deux échantillons. Pour réduire la probabilité de dessiner les amorces dans des régions divergentes entre les différentes copies potentielles, plusieurs couples pourraient être dessinés pour chaque gène réarrangé. Par la suite, pour les gènes ayant plusieurs copies identifiées via le Biomark, leur position pourra être déterminée par criblage PCR des bras de chromosomes triés. Cette expérience permettrait ainsi de savoir si certaines copies sont en position orthologue par rapport au riz ou à *Brachypodium* et si cette duplication interchromosomique est retrouvée sur les deux autres chromosomes homéologues 3A et 3D. A partir de ces résultats, une analyse plus précise pourrait être réalisée à partir des séquences des copies potentiellement à l'origine des gènes réarrangés. Les BAC portant ces gènes pourraient être récupérés après criblage PCR des banques BAC chromosome-spécifiques ou des pools 3D des MTP des cartes physiques (si disponibles) des bras de chromosome correspondants puis être séquencés. Ces séquences aideraient vraisemblablement à identifier la copie homéologue à l'origine de la duplication interchromosomique et peut-être à dater l'évènement par analyse de la divergence de séquence.

Des analyses d'expression fines de ces gènes réarrangés et des gènes à leur origine ne pourront être menées qu'une fois leur séquence respective connue. Ces séquences pourront ainsi être clairement identifiées dans le jeu de données issu de l'expérience de RNAseq. Du fait de la forte redondance potentielle de ces séquences, les données de RNAseq pourraient s'avérer difficiles à analyser. Ainsi ces séquences pourront servir de base pour dessiner des amorces spécifiques utilisables pour mener des expériences de Q-PCR. Des données d'expression précises pour ces gènes non synténiques permettront de vérifier l'hypothèse selon laquelle ces duplications auraient été conservées au sein des îlots de gènes et sélectionnées dans leur nouvelle position pour des raisons fonctionnelles (cf. article n°3).

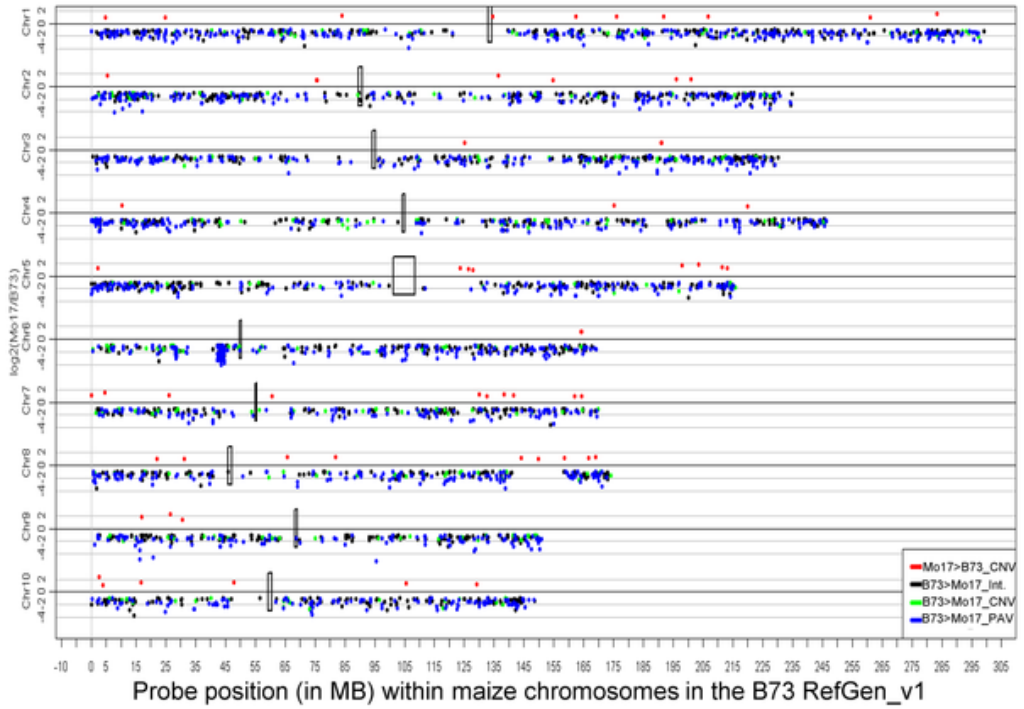


Figure 30 : Distribution des CNV et PAV dans le génome du maïs. Les dix chromosomes de maïs sont représentés. Le code couleur désigne le type de variation comme présenté dans la légende. Les positions des centromères sont représentées par des rectangles noirs. (d'après Springer et al., 2009)

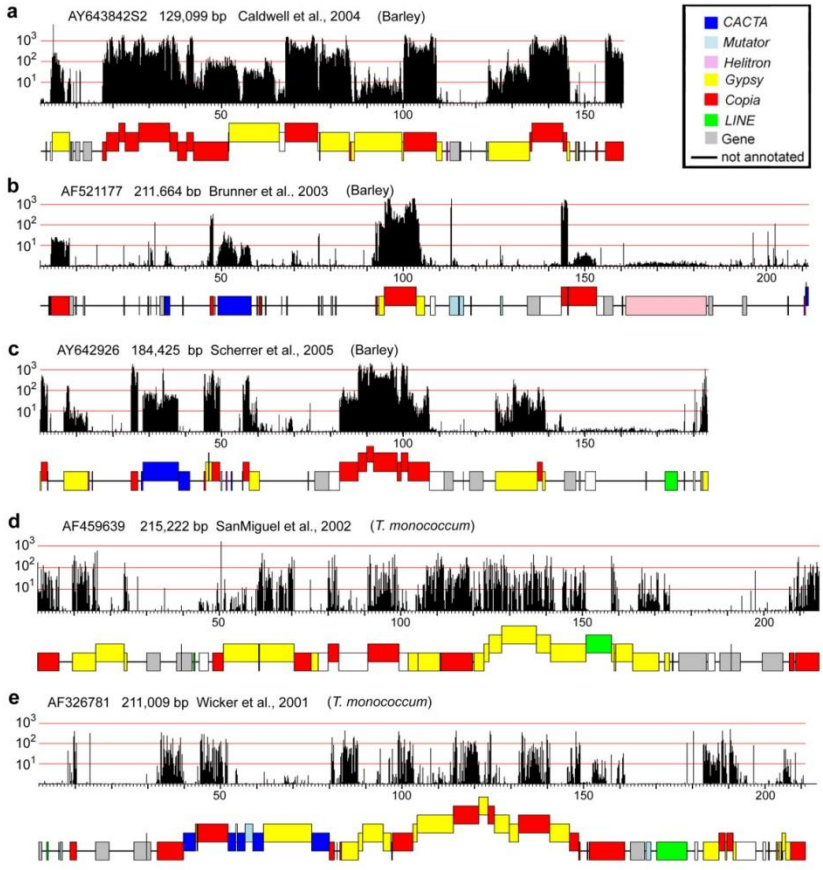


Figure 31 : Représentation graphique du MDR de plusieurs séquences et de leur annotations expertisées. Les histogrammes représentant le MDR sont situés au-dessus des schémas d'annotations expertisées des séquences. Les ET sont représentés en couleur avec une couleur par superfamille. Les ET les plus récemment insérés sont représentés au-dessus de ceux dans lesquels ils se sont insérés. Les graphiques a à c représentent des séquences d'orge alors que les graphiques d et e sont des séquences de *Triticum monococcum* pour lesquelles les valeurs de MDR sont plus faibles. (d'après Wicker et al., 2008)

3. L'ETUDE DE LA VARIATION DU CONTENU EN GENES ENTRE VARIETES ET ENTRE HOMEOLOGUES

3.1. L'étude des variations du contenu en gènes entre variétés de blé

Comme présenté précédemment, une caractéristique majeure du génome du blé est sa redondance génique d'une part du fait de la polyploïdie récente et d'autre part du fait de nombreuses duplications intra- et interchromosomiques. Ainsi, ces observations traduisent une dynamique importante de l'espace génique chez le blé. Cependant la fraction génique a toujours été considérée comme très stable et peu polymorphe chez le blé comme en témoigne le faible taux de « Single Nucleotide Polymorphism » (SNP) identifié dans les régions codantes (1 SNP toutes les 203 à 540 pb) (Somers et al., 2003; Ravel et al., 2006; Ravel et al., 2007). Ainsi, le dynamisme de l'espace génique chez le blé ne semble pas se traduire au niveau de la séquence mais peut-être au niveau des duplications de gènes.

Chez le maïs, des expériences de « Comparative Genomic Hybridisation » (CGH) ont été réalisées pour analyser la variabilité de l'espace génique à l'échelle des gènes (Figure 30) (Springer et al., 2009). Une puce représentant l'ensemble de la séquence du génome de référence du maïs a été synthétisée et hybridée conjointement avec de l'ADN de la variété de référence et celui d'une autre variété. Avec cette expérience, des variations du nombre de copies de gènes (Copy Number Variation, CNV) ou des variations du type présence/absence de certains gènes (Presence/Absence Variation, PAV) ont été caractérisées entre le génome de la variété de référence et celui d'une autre variété.

Dans le cadre du projet 3BSEQ, des expériences de CGH du même type sont planifiées. Pour ce faire, une puce dite de « tiling » (traduction : tuiles ou carrelage) sera développée à partir de la séquence du chromosome 3B. Cette puce couvrira les régions présentant de faibles valeurs d'index « Mathematically Defined Repeat » (MDR) (Kurtz et al., 2008; Wicker et al., 2008; Choulet et al., 2010), traduisant la faible redondance de ces séquences (Figure 31). Les deux brins des séquences sélectionnées seront synthétisés sur la puce sous forme d'oligonucléotides chevauchants sur au moins la moitié de leur longueur. Avec ce design, cette puce de tiling devrait permettre de mener des expériences de CGH sur l'intégralité des gènes du chromosome 3B avec une grande précision. La variabilité de l'espace génique pourra ainsi être analysée entre différentes variétés de blé à l'échelle des gènes entiers non

plus à l'échelle de la séquence comme par l'identification de SNP. Des CNV et des PAV pourront être recherchées entre les variétés choisies et la variété de référence Chinese Spring dont la séquence sera synthétisée sur la puce.

L'observation de ces variations pourra également être couplée à une analyse fonctionnelle des gènes soumis à variations entre variétés. En utilisant les profils d'expression établis pour les gènes de la variété Chinese Spring, il sera possible de voir si les gènes soumis à variation sont exprimés dans la variété de référence et donc s'ils contribuent potentiellement à son phénotype.

De plus, des études ont montré un impact du nombre de copies sur certains phénotypes chez le blé. Par exemple, l'analyse du gène Q a montré que la présence de moins de quatre copies de l'allèle q dans un génome, produisait des épis fins, élancés et peu productifs alors que la présence de plus de cinq copies engendrait des épis compacts et productifs (Faris et al., 2003). Il serait donc envisageable de lier les CNV et les PAV observées à des variations phénotypiques de certains caractères d'intérêt par association.

3.2. L'étude des variations homéologues du contenu en gènes et de leur évolution

En outre, ces expériences de CGH pourront être adaptées à l'analyse de l'évolution du contenu génique des génomes homéologues présents dans le génome du blé tendre. En effet, des études menées sur l'impact de la polyploïdie sur la structure du génome de blé ont permis d'observer une diminution de taille de 2% à 10% par rapport aux génomes progéniteurs diploïdes (Feldman and Levy, 2009). De plus, comme énoncé précédemment, le génome du blé compte également une grande proportion de gènes dupliqués indépendamment de la polyploïdie (Akhunov et al., 2003; Choulet et al., 2010). Ainsi, la recherche de CNV ou PAV sera menée non plus entre variétés différentes mais entre les chromosomes homéologues 3A, 3B et 3D qui seront hybridés individuellement sur la puce pour identifier des CNV ou PAV homéologues. Les polymorphismes géniques pourront donc être analysés afin de mieux comprendre les dynamiques de diminution et d'augmentation de la redondance génique entre chromosomes homéologues.

L'évolution de ces variations homéologues pourra également être étudiée afin de mieux comprendre l'impact de la polyploïdisation sur la redondance génique chez le blé. Une expérience de CGH sera réalisée avec les progéniteurs potentiels du blé tendre actuel tel que le blé dur (génome tétraploïde AB), *Triticum monococcum* (génome A), *Aegilops speltoides* (ancêtre potentiel du génome B) et *Aegilops tauschii* (génome D) pour comprendre les dynamiques de la redondance génique au cours de la formation du génome du blé. Des espèces proches telles que l'orge pourront également être étudiées afin de savoir si la dynamique observée chez le blé lui est spécifique ou si elle peut être généralisable aux Triticeae.

3.3. Un projet pilote sur les régions faiblement répétées des contigs séquencés

Avant de réaliser ces expériences de CGH sur les séquences faiblement répétées du chromosome 3B entier, un projet pilote sera réalisé pour valider la faisabilité et les méthodes d'analyse à partir des régions faiblement répétées des contigs séquencés du chromosome 3B (Choulet et al., 2010).

Pour ce faire, j'ai développé une puce de tiling à partir des 18 Mb de séquences annotées du chromosome 3B en sélectionnant les régions faiblement répétées. Ainsi, plus de 5 Mb de séquences faiblement répétées correspondant à des régions géniques (71%), à des régions non annotées (16%) et à des régions composées d'éléments transposables (13%) probablement présents en petit nombre de copies dans le génome ont été sélectionnées. Les deux brins des séquences sélectionnées ont été synthétisés sur la puce sous forme d'oligonucléotides de 60 nt chevauchants de 45 nt en moyenne.

Si les résultats du projet pilote s'avèrent concluants, les expériences de CGH seront renouvelés sur la puce de tiling portant les régions faiblement répétées du chromosome 3B entier. Si un design identique est retenu et si le chromosome 3B porte la même proportion de régions faiblement répétées que les contigs séquencés soit environ 28%, 280 Mb seraient sélectionnés pour être synthétisés sur cette puce. De plus, avec le développement rapide des technologies de puce permettant l'augmentation de la densité de sondes synthétisées sur les

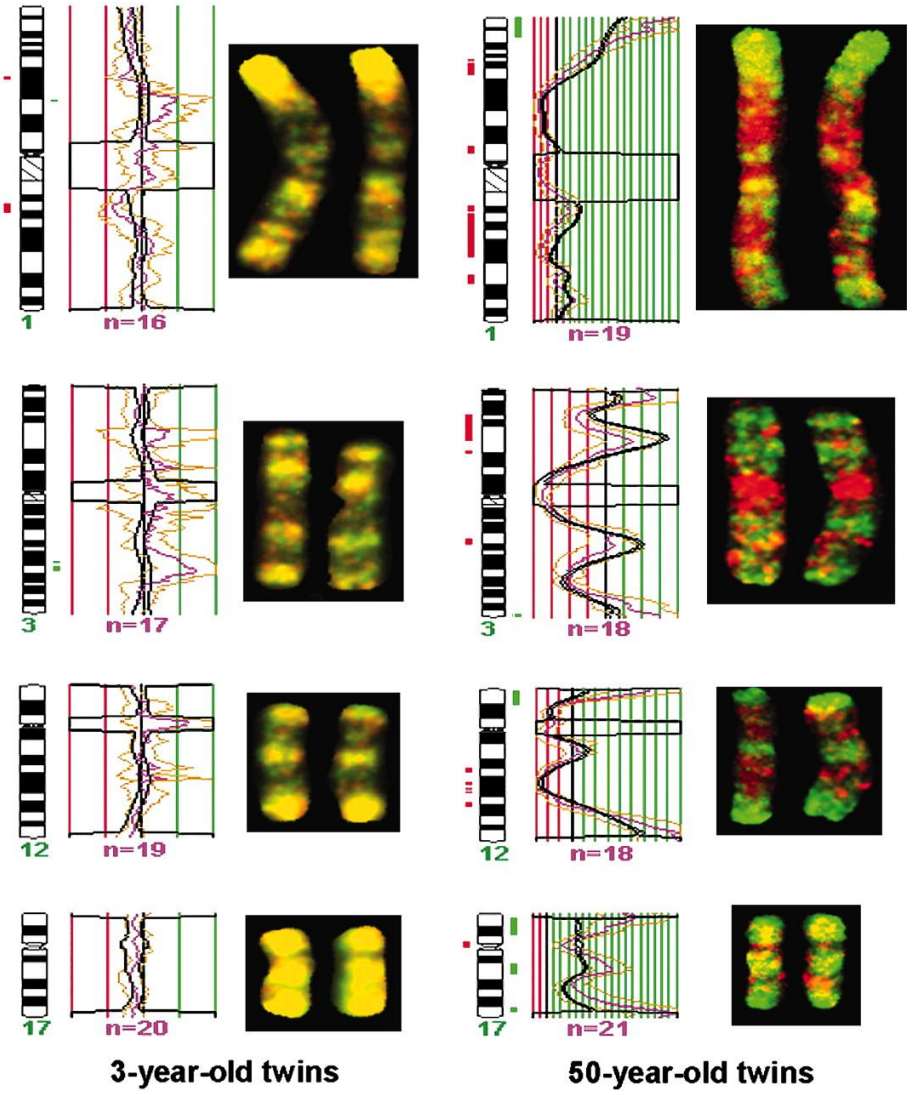


Figure 32 : Cartographie des régions chromosomiques avec des profils de méthylations d'ADN différents entre deux jumeaux monozygotiques par Digital Karyotyping des régions méthylées. Ces exemples montrent les hybridations obtenues sur les chromosomes 1, 3, 12 et 17 de deux paires de jumeaux âgés de 3 et 50 ans. Le vert indique l'hyperméthylation et le rouge l'hypométhylation. Très peu de différences significatives sont observables entre les jumeaux de 3 ans (représentées par les bandes vertes et rouges sur les côtés des chromosomes) alors que les profils sont plus différents entre les jumeaux de 50 ans. (d'après Fraga et al., 2005)

lames, l'intégralité de la séquence du chromosome 3B pourrait peut-être être portée par une puce à moyen terme.

4. L'ETUDE DES MARQUES EPIGENETIQUES ET DE LEUR IMPACT SUR L'EXPRESSION DES GENES

L'article n°3 a permis d'observer en filigrane le rôle majeur d'un mécanisme de régulation de l'expression des gènes probablement lié aux marques épigénétiques. Comme les analyses des marques épigénétiques sont essentiellement basées sur la séquence des génomes, le blé a jusqu'à présent été complètement mis à l'écart de ce type d'analyses. Cependant, la mise à disposition prochaine de la séquence du chromosome 3B permettra d'explorer un champ d'étude complètement nouveau dans l'analyse du génome du blé et d'approfondir les relations entre l'organisation de l'espace génique et les régulations des gènes par les marques épigénétiques.

4.1. Les marques épigénétiques le long du chromosome 3B

Dans un premier temps, une analyse à l'échelle du chromosome entier permettrait de vérifier le statut particulier des télomères par rapport aux autres régions du chromosome 3B. En effet, les télomères du chromosome 3B seraient enrichis en clusters de gènes coexprimés mais aussi en îlots de corégulation (Cf. article n°3). Ceci suggère donc une dynamique importante et une forte activité transcriptionnelle dans ces régions distales. Ainsi les régions télomériques sont peut-être soumises à des régulations du type épigénétiques particulières par rapport aux régions plus proximales des chromosomes. Pour vérifier cette hypothèse, une expérience de « Digital Karyotyping » pourrait être tentée afin de visualiser de façon qualitative le niveau de méthylation globale des différentes régions chromosomiques (Paz et al., 2003; Fraga et al., 2005). Cette technique repose sur l'utilisation d'enzymes de restriction sensibles à la méthylation et générant des bouts cohésifs telles que *TspMI*. Une PCR peut ensuite être réalisée après la ligation d'adaptateurs aux extrémités des fragments. Avec l'utilisation de *TspMI* par exemple, cette amplification permet de sélectionner les fragments hypométhylés qui après marquage fluorescent, peuvent être hybridés avec des chromosomes en métaphase.

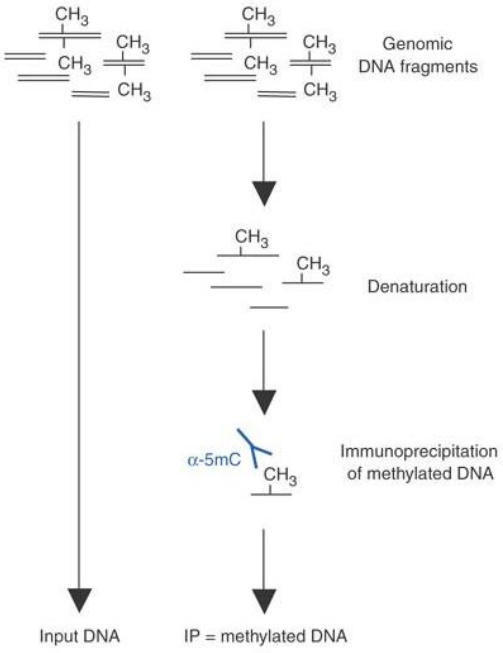


Figure 33 : Analyse de la méthylation par immunoprécipitation de l'ADN (MeDIP). (a) De l'ADN génomique dénaturé de la taille souhaitée (obtenue par digestion ou par sonication) est incubé avec un anticorps dirigé contre les 5-methyl-cytosine (α-5mC) et l'ADN méthylé est isolé par immunoprécipitation (IP). L'enrichissement en séquences cibles dans la fraction méthylée peut être quantifié par des méthodes de détection d'ADN classique comme la PCR ou les puces. (d'après Weber et al., 2005)

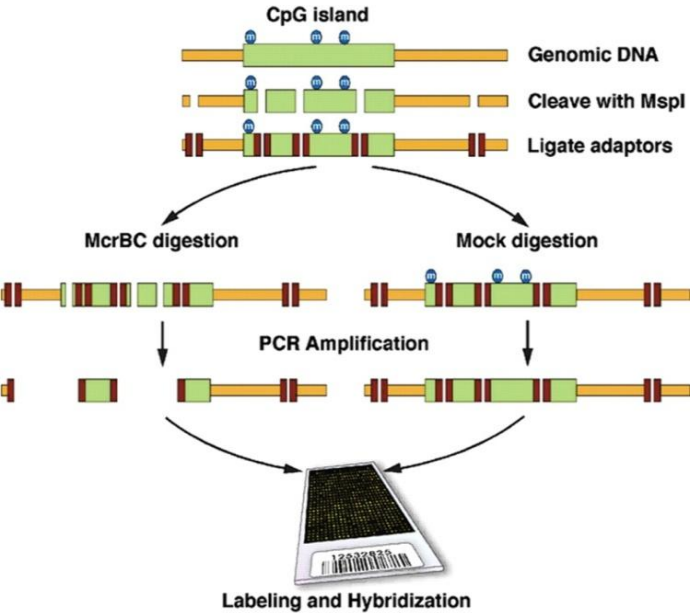


Figure 34 : Schéma du protocole « McrBC assay » sur puce. L'ADN génomique avec un îlot CpG est représenté en haut de la figure. Cet ADN est clivé par l'enzyme de restriction MspI au niveau de tous les sites CCGG méthylés ou non. Puis des adaptateurs sont liés au niveau des bouts cohésifs réalisés suite à la digestion. L'échantillon est ensuite divisé en parts égales dont une subira une digestion avec McrBC qui clive au niveau des sites méthylés et l'autre subira une digestion « mock » c'est-à-dire sans enzyme. Ces deux échantillons servent de matrice pour réaliser une amplification PCR dont les produits sont utilisés pour faire une hybridation comparée. (d'après Kamalakaran et al., 2009)

Des profils de méthylation globale des chromosomes peuvent ainsi être visualisés très simplement comme présenté sur la figure 32.

D'autres analyses plus fines pourraient également être réalisées grâce à l'utilisation de la séquence du chromosome 3B et plus particulièrement de la puce de tiling du chromosome 3B entier. Des expériences de « Methylation DNA ImmunoPrecipitation » (MeDIP) sur puce (MeDIP-chip) (Weber et al., 2005) pourraient être mises en œuvre pour établir les profils de méthylation de l'ADN le long du chromosome 3B (Figure 33). Par cette technique, l'ADN génomique préalablement fragmenté est immunoprécipité avec des anticorps orientés contre les cytosines méthylées. Après élution, un mélange de fragments d'ADN enrichis en fragments hyperméthylés est obtenu. Les fragments hyperméthylés marqués d'une couleur seraient co-hybridés avec des fragments d'ADN non enrichis sur la puce de tiling. Après traitement des données d'hybridation, les sondes correspondant aux fragments hyperméthylés seraient identifiées et repositionnées le long du chromosome pour établir le profil de méthylation fin du chromosome 3B. Cette expérience pourrait être réalisée à partir des 15 échantillons d'ADN utilisés pendant ma thèse. Ainsi le profil de méthylation établi sur ces 15 échantillons pourrait ensuite être comparé aux profils d'expression des gènes pour étudier les relations entre le niveau de méthylation et l'expression. L'inconvénient de cette technique de visualisation des fragments hyperméthylés sur puce est que la distinction des séquences homéologues sera probablement impossible et que les profils établis seront à nouveau des profils cumulés de plusieurs gènes. De même que le RNAseq est envisagé comme technique alternative aux puces d'expression, le recours au MeDIP-seq serait une alternative plus précise au MeDIP-chip. Ainsi, après séquençage à haut-débit, les séquences de fragments méthylés pourraint être identifiées de façon précise le long de la séquence du chromosome 3B.

Cependant, les fragments hyperméthylés seront probablement fortement enrichis en ET ce qui pourraient générer des données de séquences ou d'hybridation difficiles à analyser du fait de leur forte redondance. Ainsi, les techniques ciblant les régions hypométhylées seraient peut-être plus adaptées à l'analyse du méthylome du génome de blé. La technologie décrite par Kamalakaran et al. (2009) de « McrBC assay » sur puce (Figure 34) permet d'amplifier spécifiquement les fragments hypométhylés d'un génome. Cette fraction hypométhylée du génome de blé pourrait ensuite être hybridée sur la puce tiling conjointement avec l'ADN

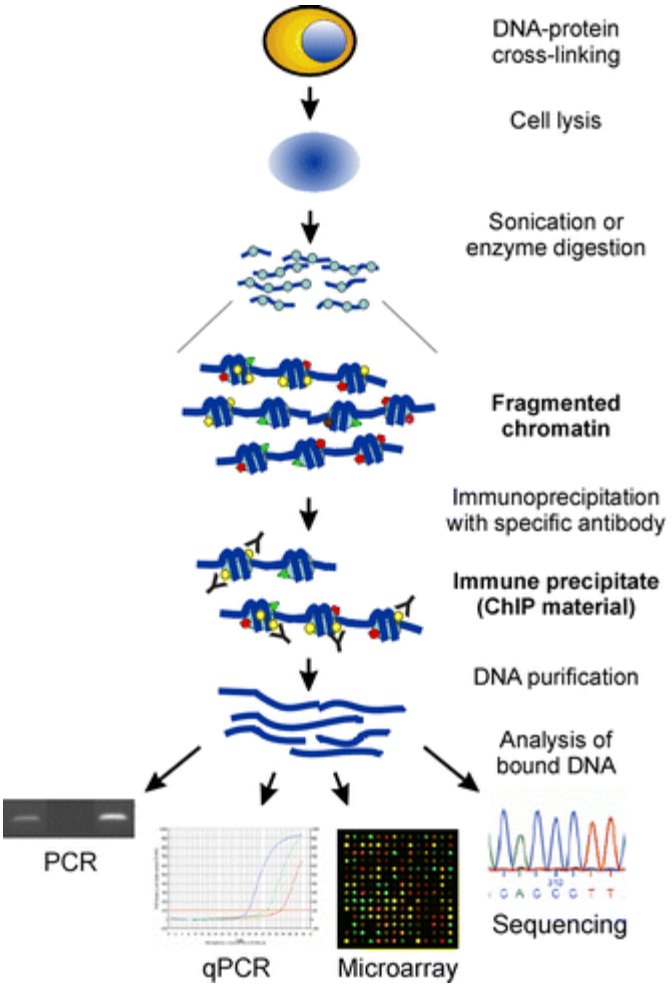


Figure 35 : L'enrichissement en fragments d'ADN associés à des modifications de la chromatine par immunoprécipitation de la chromatine (ChIP) et ses méthodes d'analyse variées. (d'après Collas, 2010)

génomique non enrichi pour identifier par comparaison les fractions hyperméthylées dans le génome.

Pour analyser la distribution des différentes marques de modifications des histones le long du chromosome 3B, une expérience basée sur l'immunoprécipitation pourrait à nouveau être réalisée comme la technique de « Chromatin ImmunoPrecipitation » (ChIP) sur puce (ChIP-chip) ou par séquençage (ChIP-seq) (Figure 35) (Collas, 2010). Après ligation des protéines à l'ADN et fragmentation, les fragments d'ADN sont immunoprécipités avec des anticorps orientés contre la modification d'histone ciblée. Les fragments d'ADN associés à cette modification d'histone peuvent ensuite être identifiés via l'hybridation sur puce ou par séquençage haut-débit. Les profils de modifications d'histones pourront également être comparés aux profils d'expression et aux profils de méthylation. Une attention particulière pourrait alors être portée à l'analyse des clusters de gènes coexprimés. En effet, l'hypothèse avancée dans l'article n°3 est que les gènes réarrangés ont été sélectionnés dans leur nouvelle position car la coexpression avec les gènes voisins apportait un avantage sélectif à la plante. Cette coexpression positivement sélectionnée serait induite par un état de la structure chromatinienne partagée entre les gènes réarrangés et les gènes voisins (Gierman et al., 2007). Ainsi, cette hypothèse pourrait être éprouvée en menant une analyse des modifications des histones dans les régions de coexpression.

La mise en place de ces expériences permettra donc d'analyser les relations entre l'organisation de l'espace génique et la régulation des gènes via les marques épigénétiques au niveau des régions faiblement répétées du chromosome 3B dans un premier temps via la puce de tiling puis peut-être à l'échelle plus fine et plus résolutive du chromosome 3B entier via le séquençage en haut-débit.

4.2. Les analyses ciblées des marques épigénétiques

Des analyses plus précises de certaines régions particulières seraient également envisageables. Comme suggéré précédemment, la répression épigénétique d'un ET à proximité d'un gène peut induire la répression de l'expression de ce gène (Hollister and Gaut, 2009). Le génome du blé, étant riche en ET, serait un excellent modèle d'étude pour explorer l'impact des ET sur l'expression des gènes dans leur entourage direct. Les méthodes de MeDIP-chip ou de

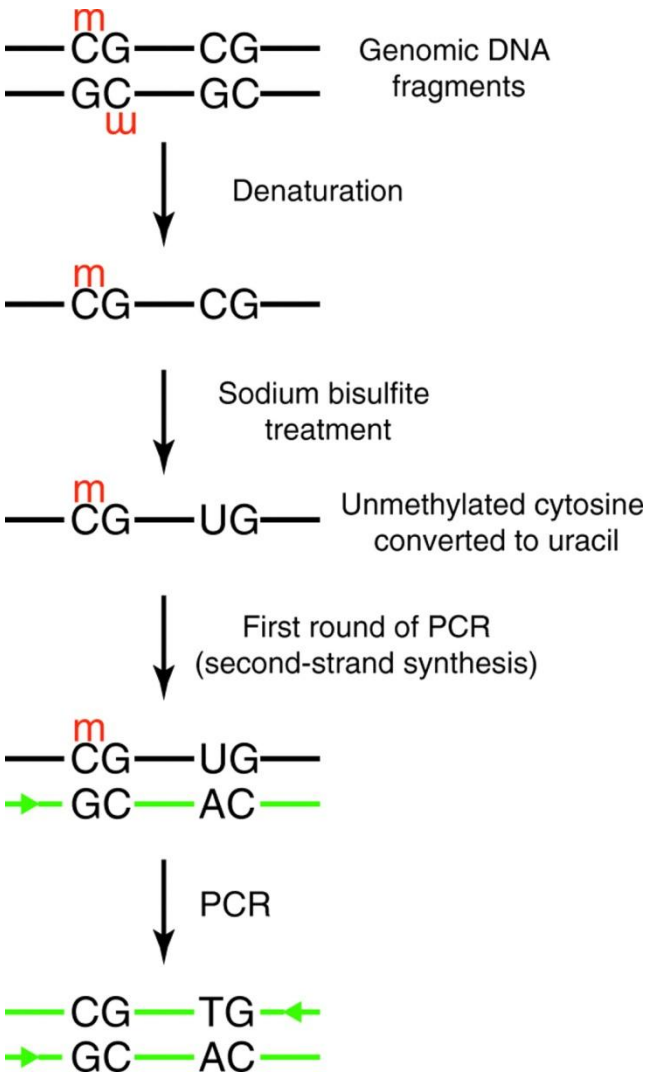


Figure 36 : La conversion bisulfite. L'ADN est dénaturé puis traité avec du bisulfite de sodium pour convertir les cytosines non méthylées en uraciles, qui sont convertis en thymine par PCR. (d'après Zilberman et Henikoff, 2007)

ChIP-chip ne permettraient probablement pas d'aborder cette question de façon satisfaisante car les ET sont présents en nombre de copies important dans le génome du blé. Ainsi une sonde correspondant à un ET serait hybridée avec des fragments d'ET issus de multiples locus dans le génome. Le même type de problème liée à la redondance des séquences d'ET sera observé avec les techniques basées sur le séquençage à haut-débit telles que le MeDIP-seq ou le ChIP-seq. Les séquences de petite taille rendront difficile leur assignation à un emplacement précis dans le génome.

Ainsi, une approche plus ciblée à plus faible débit serait vraisemblablement plus adaptée. Par exemple, grâce à la séquence du chromosome 3B, des marqueurs du type « Insertion Site-Based Polymorphism » (ISBP; Paux et al., 2010) correspondant à des insertions d'ET dans une autre séquence, ET d'une autre famille ou région faiblement répétée, pourraient être dessinés. Comme ces insertions sont généralement uniques dans le génome, ces marqueurs permettraient de cibler des séquences d'ET dans le génome pour étudier leur niveau de méthylation ou leur association avec des types de modifications d'histone particulières. Ces marqueurs, utilisés en Q-PCR sur les mélanges de fragments d'ADN enrichis en fragments méthylés (MeDIP) ou associés à une modification d'histone (ChIP), permettraient de quantifier les marques épigénétiques mises en place au niveau de la séquence d'ET et d'étudier leur impact sur la régulation de l'expression des gènes à proximité. Une alternative aux techniques utilisant l'immunoprécipitation serait l'utilisation de la technique du « bisulfite sequencing » pour analyser les méthylations des cytosines. En effet, le traitement au bisulfite d'ADN génomique transforme les cytosines non méthylées en uracile (Figure 36) (Zilberman and Henikoff, 2007). Ainsi, la comparaison d'une séquence d'ISBP traité au bisulfite avec la séquence de référence permettrait de connaître avec précision la position des cytosines méthylées au sein de l'ET.

Ces approches plus ciblées pourraient également être utilisées pour analyser les régulations mises en place au niveau d'autres séquences dont la redondance compliquerait les analyses via les techniques à haut-débit. Par exemple, l'analyse des marques épigénétiques portées par des gènes homéologues permettrait de comprendre les mécanismes de répression mis en place pour contrer les effets néfastes de la polyploïdie sur le fonctionnement du génome. De plus, une analyse évolutive de ces marques pourrait être réalisée en étudiant également les génomes diploïdes ancêtres du génome de blé. Le même type d'analyse pourrait être mis en œuvre avec les gènes dupliqués dans le génome du blé et plus particulièrement les gènes réarrangés pour

mieux comprendre les mécanismes mis en place pour réguler la redondance d'informations géniques. Il serait ainsi possible d'observer et de comparer les marques épigénétiques portées par le gène d'origine avec celles portées par le gène dupliqué.

En conclusion, une nouvelle thématique de recherche complètement inexplorée chez le blé est en train d'émerger grâce à la séquence du chromosome 3B. Elle ouvre de toutes nouvelles perspectives sur la compréhension du fonctionnement du génome du blé et notamment sur les relations entre l'organisation de l'espace génique et les régulations épigénétiques des gènes et des ET.

4.3. Les analyses des régulations à longue distance impliquées dans les îlots de corégulation

Un autre champ thématique complètement vierge de toutes études chez le blé est l'analyse des territoires chromosomiques et des régulations à longue distance impliquant probablement des interactions intra- et interchromosomiques. Dans l'article n°3, des îlots de corégulation au sein du chromosome 3B ont été mis en évidence essentiellement au niveau des télomères. De plus, des études ont montré que les chromosomes de blé adoptaient la configuration Rabl à l'interphase, à savoir que les chromosomes restaient partiellement condensés, que les centromères étaient localisés à une extrémité de la cellule et les télomères à l'opposé et que les bras du même chromosome restaient proches (Dong and Jiang, 1998; Cowan et al., 2001; Santos et al., 2002). Ainsi, nous avons émis l'hypothèse que cette configuration devait favoriser les interactions intrachromosomiques à longue distance. Pour vérifier cette hypothèse basée sur la configuration Rabl, une expérience de « Fluorescence In Situ Hybridization » (FISH) pourrait être réalisée. Comme des séquences télomériques ont été identifiées chez le blé (Mao et al., 1997), leur marquage suivi de leur hybridation sur des noyaux cellulaires en interphase permettrait de visualiser la position des télomères « décondensés » pour vérifier s'ils sont maintenus proches à une extrémité des cellules.

Les îlots de corégulation pourraient également être étudiés plus en détail pour savoir si leur mécanisme de régulation commun implique une proximité des clusters de gènes coexprimés dans les noyaux cellulaires en interphase lorsque les chromosomes sont décondensés. La visualisation de ce qui a pu être appelé « gene kissing » (Lanctôt et al., 2007) a été notamment

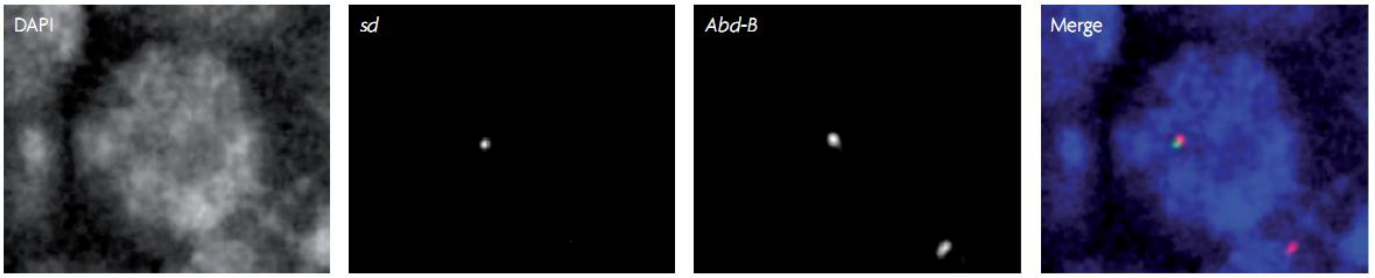


Figure 37 : « Gene kissing ». Dans cet exemple de « gene kissing » chez la Drosophile, deux locus impliqués dans la régulation par Fab7 qui sont localisés sur des chromosomes différents, colocalisent dans le noyau des cellules. Le DAPI est utilisé pour marquer les noyaux, *sd* marque la position d'une copie transgénique de Fab7 insérée dans le chromosome X au niveau du locus *scalloped* (*sd*). *Abd-B* montre le locus régulé par l'élément Fab7. Les deux locus colocalisent (« kiss ») dans une fraction significative des noyaux comme sur l'image de droite où toutes les images sont superposées. (d'après Lanctôt et al., 2007)

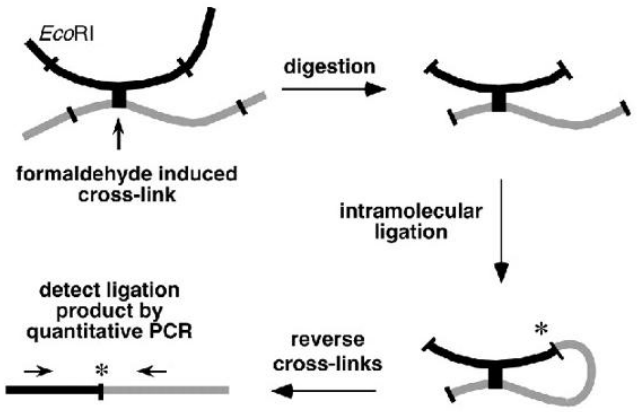


Figure 38 : La méthodologie « Chromosome Conformation Capture » (3C). L'expérience est menée comme suit : formation des liaisons avec du formaldehyde, digestion EcoRI, ligation intramoléculaire et détection des produits de ligation par PCR après suppression des liaisons aux protéines. (d'après Dekker et al., 2002)

réalisée chez la Drosophile par FISH avec un marquage des fragments supposés interagir en deux couleurs différentes (Figure 37). Le FISH est une technique éprouvée chez le blé mais nécessite des fragments de grande taille pour générer une hybridation spécifique visible. Ainsi, des BAC sont généralement utilisés en FISH chez le blé. Le marquage et l'hybridation des BAC portant les clusters de coexpression supposés interagir pourraient être tentés mais la présence d'ET dans ces BAC pourraient générer des signaux diffus sur des chromosomes décondensés. Une alternative à l'utilisation de BAC de blé serait d'avoir recours à des BAC de riz ou de *Brachypodium* si les clusters de gènes sont constitués de gènes colinéaires entre ces espèces. En effet, les ET du riz ou de *Brachypodium* seraient peut-être suffisamment différents de ceux présents dans le génome du blé pour ne pas générer de signaux aspécifiques et ainsi mieux marquer les régions d'intérêt. Cependant comme ces visualisations d'interactions sont réalisées en deux dimensions, un grand nombre d'expériences seraient nécessaires pour valider statistiquement une interaction entre les BAC. Le recours à la technologie du FISH en trois dimensions (3D-FISH; Tirichine et al., 2009) permettrait de contourner cette limite. Ainsi une collaboration avec l'équipe dirigée par Chantal Vaury de l'UMR « Génétique Reproduction & Développement » de Clermont-Ferrand maîtrisant cette technologie pourrait être envisagée pour visualiser les interactions de gènes à longue distance et en trois dimensions.

Ces deux approches basées sur la technique du FISH permettraient de mettre en évidence une organisation tridimensionnelle de chromosomes décondensés dans le noyau et éventuellement des interactions entre régions chromosomiques distantes. Ces analyses cytogénétiques de visualisation pourraient être complétées par une expérience de « Chromosome Conformation Capture » (3C) (Figure 38) (Dekker et al., 2002). La technique 3C est utilisée pour déterminer quelles séquences d'ADN sont proches les unes des autres dans l'espace tridimensionnel du noyau de cellules fixées. Des liaisons covalentes sont créées au niveau des sites d'interactions du type ADN-protéine-ADN grâce à un traitement au formaldéhyde. Les séquences d'ADN proches les unes des autres sont ainsi liées les unes aux autres. L'ADN est ensuite digéré avec une enzyme de restriction puis soumis à une ligation à faible concentration. Dans de telles conditions, les liaisons intramoléculaires sont favorisées par rapport aux liaisons entre molécules différentes. Les liaisons ADN-protéine-ADN sont ensuite supprimées et les séquences d'ADN initialement proches dans le noyau sont ainsi liées par leurs extrémités. Une Q-PCR peut ensuite être réalisée avec une amorce dans chacune des séquences

suspectées interagir au sein du noyau afin d'évaluer la fréquence d'interaction de ces séquences. Cette technique pourrait à nouveau être mise en œuvre chez le blé en collaboration avec l'équipe dirigée par Chantal Vaury maîtrisant également cette technologie.

En conclusion, mon travail de thèse a permis de confirmer des hypothèses et d'obtenir des résultats originaux concernant l'impact de la structure du génome sur l'organisation, la régulation et la fonction des gènes sur le chromosome 3B du blé tendre. Il a également permis de formuler de nombreuses hypothèses nouvelles quant à la formation et au maintien différentiels des îlots de gènes le long du chromosome 3B. Ainsi, ma thèse débouche sur des perspectives variées et originales qui, je l'espère, mèneront vers une connaissance de plus en plus approfondie du fonctionnement de ce génome si complexe et qui aideront indirectement à l'amélioration de cette espèce végétale d'importance majeure qu'est le blé tendre.

LISTE DES REFERENCES BIBLIOGRAPHIQUES

Abranches, R., Beven, A.F., Aragon-Alcaide, L., and Shaw, P.J. (1998). Transcription sites are not correlated with chromosome territories in wheat nuclei. J. Cell Biol. **143:** 5-12.

Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T., Tannier, E., Plomion, C., Cooke, R., Feuillet, C., et al. (2010). Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends Plant Sci. **15:** 479-487.

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

Akhunov, E.D., Goodyear, A.W., Geng, S., Qi, L.L., Echalier, B., Gill, B.S., Miftahudin, Gustafson, J.P., Lazo, G., Chao, S., et al. (2003). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. **13:** 753-763.

Akhunov, E.D., Akhunova, A.R., and Dvorak, J. (2007). Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. Mol. Biol. Evol. **24:** 539-550.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389-3402.

Animal genome size database, A. (2010). http://www.genomesize.com.

Babu, M.M., Janga, S.C., de Santiago, I., and Pombo, A. (2008). Eukaryotic gene regulation in three dimensions and its impact on genome evolution. Curr. Opin. Genet. Dev. **18:** 571-582.

Batada, N.N., Urrutia, A.O., and Hurst, L.D. (2007). Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. **23:** 480-484.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., Sanmiguel, P.J., and Bennetzen, J.L. (2009a). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. **5:** e1000732.

Baucom, R.S., Estill, J.C., Leebens-Mack, J., and Bennetzen, J.L. (2009b). Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. **19:** 243-254.

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. (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. **42:** 251-269.

Bennetzen, J.L., and Ramakrishna, W. (2002). Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol. Biol. **48:** 821-827.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Ann. Bot. **95:** 127-132.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science **306:** 2242-2246.

Bertone, P., Gerstein, M., and Snyder, M. (2005). Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res. **13:** 259-274.

Bilgic, H., Cho, S., Garvin, D.F., and Muehlbauer, G.J. (2007). Mapping barley genes to chromosome arms by transcript profiling of wheat-barley ditelosomic chromosome addition lines. Genome **50:** 898-906.

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

Bossolini, E., Wicker, T., Knobel, P.A., and Keller, B. (2007). Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J. **49:** 704-717.

Bowers, J.E., Arias, M.A., Asher, R., Avise, J.A., Ball, R.T., Brewer, G.A., Buss, R.W., Chen, A.H., Edwards, T.M., Estill, J.C., et al. (2005). Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc. Natl. Acad. Sci. U. S. A. **102:** 13206-13211.

Breen, J., Wicker, T., Kong, X., Zhang, J., Ma, W., Paux, E., Feuillet, C., Appels, R., and Bellgard, M. (2010). A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. BMC Plant Biol. **10:** 98.

Brooks, S.A., Huang, L., Gill, B.S., and Fellers, J.P. (2002). Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. Genome **45:** 963-972.

Carpentier, A.S., Torresani, B., Grossmann, A., and Henaut, A. (2005). Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data. BMC Genomics **6:** 84.

Casacuberta, J.M., and Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene **311:** 1-11.

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.

Chan, A.P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K.M., Redman, J., Chen, G., et al. (2010). Draft genome sequence of the oilseed species Ricinus communis. Nat. Biotechnol. **28:** 951-956.

Chantret, N., Cenci, A., Sabot, F., Anderson, O., and Dubcovsky, J. (2004). Sequencing of the Triticum monococcum hardness locus reveals good microcolinearity with rice. Mol. Genet. Genomics **271:** 377-386.

Chantret, N., Salse, J., Sabot, F., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.F., et al. (2008). Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. J. Mol. Evol. **66:** 138-150.

Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., Segurens, B., Carter, M., Huteau, V., Coriton, O., et al. (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics **180:** 1071-1086.

Chen, W.H., de Meaux, J., and Lercher, M.J. (2010a). Co-expression of neighbouring genes in Arabidopsis: separating chromatin effects from direct interactions. BMC Genomics **11:** 178.

Chen, X., Hackett, C.A., Niks, R.E., Hedley, P.E., Booth, C., Druka, A., Marcel, T.C., Vels, A., Bayer, M., Milne, I., et al. (2010b). An eQTL analysis of partial resistance to Puccinia hordei in barley. PLoS One **5:** e8598.

Cho, S., Garvin, D.F., and Muehlbauer, G.J. (2006). Transcriptome analysis and physical mapping of barley genes in wheat-barley chromosome addition lines. Genetics **172:** 1277-1285.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.C., Magdelenat, G., Gonthier, C., et al. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell **22:** 1686-1701.

Cloonan, N., and Grimmond, S.M. (2008). Transcriptome content and dynamics at single-nucleotide resolution. Genome Biol. **9:** 234.

Close, T.J., Bhat, P.R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J.T., Wanamaker, S., et al. (2009). Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics **10:** 582.

Clough, S.J., Tuteja, J.H., Li, M., Marek, L.F., Shoemaker, R.C., and Vodkin, L.O. (2004). Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the I locus. Genome **47:** 819-831.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. **26:** 183-186.

Collas, P. (2010). The current state of chromatin immunoprecipitation. Mol. Biotechnol. **45:** 87-100.

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

Cone, K.C., McMullen, M.D., Bi, I.V., Davis, G.L., Yim, Y.S., Gardiner, J.M., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Schroeder, S.G., et al. (2002). Genetic, physical, and informatics resources for maize. On the road to an integrated map. Plant Physiol. **130:** 1598-1605.

Conley, A.B., Miller, W.J., and Jordan, I.K. (2008). Human cis natural antisense transcripts initiated by transposable elements. Trends Genet. **24:** 53-56.

Conley, E.J., Nduati, V., Gonzalez-Hernandez, J.L., Mesfin, A., Trudeau-Spanjers, M., Chao, S., Lazo, G.R., Hummel, D.D., Anderson, O.D., Qi, L.L., et al. (2004). A 2600-locus chromosome bin map of wheat homoeologous group 2 reveals interstitial gene-rich islands and colinearity with rice. Genetics **168:** 625-637.

Cowan, C.R., Carlton, P.M., and Cande, W.Z. (2001). The Polar Arrangement of Telomeres in Interphase and Meiosis. Rabl Organization and the Bouquet. Plant Physiol. **125:** 532-538.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science **295:** 1306-1311.

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

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. **12:** 1075-1079.

Devos, K.M. (2005). Updating the 'crop circle'. Curr. Opin. Plant Biol. **8:** 155-162.

Devos, K.M., Ma, J., Pontaroli, A.C., Pratt, L.H., and Bennetzen, J.L. (2005). Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc. Natl. Acad. Sci. U. S. A. **102:** 19243-19248.

Devos, K.M. (2010). Grass genome organization and evolution. Curr. Opin. Plant Biol. **13:** 139-145.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry A **51:** 127-128; author reply 129.

Dong, F., and Jiang, J. (1998). Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. Chromosome Res. **6:** 551-558.

Dooner, H.K., and Weil, C.F. (2007). Give-and-take: interactions between DNA transposons and their host plant genomes. Curr. Opin. Genet. Dev. **17:** 486-492.

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., Luo, M.C., Zhong, G.Y., Bransteitter, R., Desai, A., Kilian, A., Kleinhofs, A., and Dvorak, J. (1996). Genetic map of diploid wheat, Triticum monococcum L, and its comparison with maps of Hordeum vulgare L. Genetics **143:** 983-999.

Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. Science **316:** 1862-1866.

Ducreux, L.J., Morris, W.L., Prosser, I.M., Morris, J.A., Beale, M.H., Wright, F., Shepherd, T., Bryan, G.J., Hedley, P.E., and Taylor, M.A. (2008). Expression profiling of potato germplasm differentiated in quality traits leads to the identification of candidate flavour and texture genes. J. Exp. Bot. **59:** 4219-4231.

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

Eichler, E.E., and Sankoff, D. (2003). Structural Dynamics of Eukaryotic Chromosome Evolution. Science **301:** 793-797.

Endo, T.R. (1988). Induction of chromosomal structural changes by a chromosome of Aegilops cylindrica L. in common wheat. J. Hered. **79:** 366-370.

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

Erayman, M., Sandhu, D., Sidhu, D., Dilbirligi, M., Baenziger, P.S., and Gill, K.S. (2004). Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res. **32:** 3546-3565.

Evans, A. (2009). The Feeding of the Nine Billion: Global Food Security for the 21st Century (London: Chatham House).

FAOSTAT. (2010). http://faostat.fao.org/ (Food and Agriculture Organization of the United Nations).

Faris, J.D., Fellers, J.P., Brooks, S.A., and Gill, B.S. (2003). A Bacterial Artificial Chromosome Contig Spanning the Major Domestication Locus Q in Wheat and Identification of a Candidate Gene. Genetics **164:** 311-321.

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.

Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3:** 329-341.

Feuillet, C., and Keller, B. (1999). High gene density is conserved at syntenic loci of small and large grass genomes. Proc. Natl. Acad. Sci. U. S. A. **96:** 8265-8270.

Fiston-Lavier, A.S., Anxolabehere, D., and Quesneville, H. (2007). A model of segmental duplication formation in Drosophila melanogaster. Genome Res. **17:** 1458-1470.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. U. S. A. **102:** 10604-10609.

Gai, X., and Voytas, D.F. (1998). A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. Mol. Cell **1:** 1051-1055.

Gao, X., Hou, Y., Ebina, H., Levin, H.L., and Voytas, D.F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. **18:** 359-369.

Gaut, B.S. (2002). Evolutionnary dynamics of grass genomes. New Phytol. **154:** 15-28.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res. **17:** 669-681.

Gierman, H.J., Indemans, M.H., Koster, J., Goetze, S., Seppen, J., Geerts, D., van Driel, R., and Versteeg, R. (2007). Domain-wide regulation of gene expression in the human genome. Genome Res. **17:** 1286-1295.

Gill, B.S., Appels, R., Botha-Oberholster, A.-M., Buell, C.R., Bennetzen, J.L., Chalhoub, B., Chumley, F., Dvorak, J., Iwanaga, M., Keller, B., et al. (2004). A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. Genetics **168:** 1087-1096.

Gill, K.S., Gill, B.S., Endo, T.R., and Boyko, E.V. (1996a). Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics **143:** 1001-1012.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. (1996b). Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics **144:** 1883-1891.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., and Toulmin, C. (2010). Food Security: The Challenge of Feeding 9 Billion People. Science **327:** 812-818.

Gregory, B.D., Yazaki, J., and Ecker, J.R. (2008). Utilizing tiling microarrays for whole-genome analysis in plants. Plant J. **53:** 636-644.

Gregory, T.R. (2005). The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann. Bot. **95:** 133-146.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., and Bennett, M.D. (2007). Eukaryotic genome size databases. Nucleic Acids Res. **35:** D332-338.

Greilhuber, J., Dolezel, J., Lysak, M.A., and Bennett, M.D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann. Bot. **95:** 255-260.

Grover, C.E., and Wendel, J.F. (2010). Recent Insights into Mechanisms of Genome Size Change in Plants. J. Bot. **2010:** 1-8.

Guo, W., Cai, C., Wang, C., Zhao, L., Wang, L., and Zhang, T. (2008). A preliminary analysis of genome structure and composition in Gossypium hirsutum. BMC Genomics **9:** 314.

Hanada, K., Vallejo, V., Nobuta, K., Slotkin, R.K., Lisch, D., Meyers, B.C., Shiu, S.H., and Jiang, N. (2009). The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell **21:** 25-38.

Hawkins, J.S., Grover, C.E., and Wendel, J.F. (2008). Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. Plant Science **174:** 557-562.

Helleday, T. (2003). Pathways for mitotic homologous recombination in mammalian cells. Mutat. Res. **532:** 103-115.

Hollister, J.D., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. **19:** 1419-1428.

Hossain, K.G., Kalavacharla, V., Lazo, G.R., Hegstad, J., Wentz, M.J., Kianian, P.M.A., Simons, K., Gehlhar, S., Rust, J.L., Syamala, R.R., et al. (2004). A Chromosome Bin Map of 2148 Expressed Sequence Tag Loci of Wheat Homoeologous Group 7. Genetics **168:** 687-699.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., et al. (2009). The genome of the cucumber, Cucumis sativus L. Nat. Genet. **41:** 1275-1281.

Hufton, A.L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. Curr. Opin. Genet. Dev. **19:** 600-606.

Hurst, L.D., Pal, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5:** 299-310.

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

Jackson, S., Hass Jacobus, B., and Pagel, J. (2004). The Gene Space of the Soybean Genome. In Legume Crop Genomics (AOCS Publishing).

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463-467.

Jiao, Y., Jia, P., Wang, X., Su, N., Yu, S., Zhang, D., Ma, L., Feng, Q., Jin, Z., Li, L., et al. (2005). A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. Plant Cell **17:** 1641-1657.

Jin, H., Vacic, V., Girke, T., Lonardi, S., and Zhu, J.K. (2008). Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. BMC Mol. Biol. **9:** 6.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A.H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. U. S. A. **97:** 6603-6607.

Kamalakaran, S., Kendall, J., Zhao, X., Tang, C., Khan, S., Ravi, K., Auletta, T., Riggs, M., Wang, Y., Helland, A., et al. (2009). Methylation detection oligonucleotide microarray analysis: a high-resolution method for detection of CpG island methylation. Nucleic Acids Res. **37:** e89.

Kantidze, O.L., and Razin, S.V. (2009). Chromatin loops, illegitimate recombination, and genome evolution. Bioessays **31:** 278-286.

Kapitonov, V.V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. **23:** 521-529.

Kashkush, K., and Khasdan, V. (2007). Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. Genetics **177:** 1975-1985.

Kass, E.M., and Jasin, M. (2010). Collaboration and competition between DNA double-strand break repair pathways. FEBS Lett. **584:** 3703-3708.

Kawaura, K., Mochida, K., Enju, A., Totoki, Y., Toyoda, A., Sakaki, Y., Kai, C., Kawai, J., Hayashizaki, Y., Seki, M., et al. (2009). Assessment of adaptive evolution between wheat and rice as deduced from full-length common wheat cDNA sequence data and expression patterns. BMC Genomics **10:** 271.

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

Kidwell, M.G., and Lisch, D.R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int. J. Org. Evolution **55:** 1-24.

Klein, P.E., Klein, R.R., Cartinhour, S.W., Ulanch, P.E., Dong, J., Obert, J.A., Morishige, D.T., Schlueter, S.D., Childs, K.L., Ale, M., et al. (2000). A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res. **10:** 789-807.

Koren, A., Ben-Aroya, S., and Kupiec, M. (2002). Control of meiotic recombination initiation: a role for the environment? Curr. Genet. **42:** 129-139.

Krom, N., and Ramakrishna, W. (2008). Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and populus. Plant Physiol. **147:** 1763-1773.

Kronmiller, B.A., and Wise, R.P. (2008). TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol. **146:** 45-59.

Kronmiller, B.A., and Wise, R.P. (2009). Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the rf1-associated region of maize. Plant Physiol. **151:** 483-495.

Kullman, B., Tamm, H., and Kullman, K. (2005). Fungal Genome Size Database, http://www.zbi.ee/fungal-genomesize.

Kumar, A., and Bennetzen, J.L. (1999). Plant retrotransposons. Annu. Rev. Genet. **33:** 479-532.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics **9:** 517.

La Rota, M., and Sorrells, M.E. (2004). Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. Funct. Integr. Genomics **4:** 34-46.

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat. Rev. Genet. **8:** 104-115.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics **23:** 2947-2948.

Lazo, G.R., Chao, S., Hummel, D.D., Edwards, H., Crossman, C.C., Lui, N., Matthews, D.E., Carollo, V.L., Hane, D.L., You, F.M., et al. (2004). Development of an Expressed Sequence Tag (EST) Resource for Wheat (Triticum aestivum L.): EST Generation, Unigene Analysis, Probe Selection and Bioinformatics for a 16,000-Locus Bin-Delineated Map. Genetics **168:** 585-593.

Lee, J.M., and Sonnhammer, E.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. **13:** 875-882.

Lehmensiek, A., Bovill, W., Wenzl, P., Langridge, P., and Appels, R. (2009). Genetic Mapping in the Triticeae. In Genetics and Genomics of the Triticeae, C. Feuillet and G. Muelhlbauer, eds (Berlin: Springer), pp. 201-235.

Leitch, A.R., and Leitch, I.J. (2008). Genomic Plasticity and the Diversity of Polyploid Plants. Science **320:** 481-483.

Leitch, I.J., Soltis, D.E., Soltis, P.S., and Bennett, M.D. (2005). Evolution of DNA Amounts Across Land Plants (Embryophyta). Ann. Bot. **95:** 207-217.

Leitch, I.J., Beaulieu, J.M., Chase, M.W., Leitch, A.R., and Fay, M.F. (2010). Genome Size Dynamics and Evolution in Monocots. J. Bot. **2010:** 1-18.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31:** 180-183.

Lercher, M.J., and Hurst, L.D. (2006). Co-expressed yeast genes cluster over a long range but are not regularly spaced. J. Mol. Biol. **359:** 825-831.

Letowski, J., Brousseau, R., and Masson, L. (2004). Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J. Microbiol. Methods **57:** 269-278.

Li, L., Wang, X., Sasidharan, R., Stolc, V., Deng, W., He, H., Korbel, J., Chen, X., Tongprasit, W., Ronald, P., et al. (2007). Global identification and characterization of transcriptionally active regions in the rice genome. PLoS ONE **2:** e294.

Linkiewicz, A.M., Qi, L.L., Gill, B.S., Ratnasiri, A., Echalier, B., Chao, S., Lazo, G.R., Hummel, D.D., Anderson, O.D., Akhunov, E.D., et al. (2004). A 2500-Locus Bin Map of Wheat Homoeologous Group 5 Provides Insights on Gene Distribution and Colinearity With Rice. Genetics **168:** 665-676.

Lisch, D. (2009). Epigenetic Regulation of Transposable Elements in Plants. Annu. Rev. Plant Biol. **60:** 43-66.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133:** 523-536.

Liu, R., Vitte, C.m., Ma, J., Mahama, A.A., Dhliwayo, T., Lee, M., and Bennetzen, J.L. (2007). A GeneTrek analysis of the maize genome. Proc. Natl. Acad. Sci. U. S. A. **104:** 11844-11849.

Liu, S., Yeh, C.T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D., and Schnable, P.S. (2009). Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. **5:** e1000733.

Liu, X., and Han, B. (2009). Evolutionary conservation of neighbouring gene pairs in plants. Gene **437:** 71-79.

Lomiento, M., Jiang, Z., D'Addabbo, P., Eichler, E.E., and Rocchi, M. (2008). Evolutionary-new centromeres preferentially emerge within gene deserts. Genome Biol. **9:** R173.

Ma, J., Devos, K.M., and Bennetzen, J.L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. **14:** 860-869.

Mao, L., Devos, K.M., Zhu, L., and Gale, M.D. (1997). Cloning and genetic mapping of wheat telomere-associated sequences. Mol. Gen. Genet. **254:** 584-591.

Matsumoto, T., Wu, J.Z., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B.A., Baba, T., et al. (2005). The map-based sequence of the rice genome. Nature **436:** 793-800.

Mayer, K.F., Taudien, S., Martis, M., Simkova, H., Suchankova, P., Gundlach, H., Wicker, T., Petzold, A., Felder, M., Steuernagel, B., et al. (2009). Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. **151:** 496-505.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245-250.

Miftahudin, Ross, K., Ma, X.-F., Mahmoud, A.A., Layton, J., Milla, M.A.R., Chikmawati, T., Ramalingam, J., Feril, O., Pathan, M.S., et al. (2004). Analysis of

Expressed Sequence Tag Loci on Wheat Chromosome Group 4. Genetics **168:** 651-663.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L.T., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature **452:** 991-996.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in Arabidopsis. Nature **461:** 427-430.

Moore, G., Devos, K.M., Wang, Z., and Gale, M.D. (1995). Cereal genome evolution. Grasses, line up and form a circle. Curr. Biol. **5:** 737-739.

Munkvold, J.D., Greene, R.A., Bermudez-Kandianis, C.E., La Rota, C.M., Edwards, H., Sorrells, S.F., Dake, T., Benscher, D., Kantety, R., Linkiewicz, A.M., et al. (2004). Group 3 Chromosome Bin Maps of Wheat and Their Relationship to Rice Chromosome 1. Genetics **168:** 639-650.

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., and Tanisaka, T. (2008). A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. Genes Genet. Syst. **83:** 321-329.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

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.

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., et al. (2008). A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. Science **322:** 101-104.

Paux, E., and Sourdille, P. (2009). A Toolbox for Triticeae Genomics. In Genetics and Genomics of the Triticeae, C. Feuillet and G. Muelhlbauer, eds (Berlin: Springer), pp. 255-283.

Paux, E., Faure, S., Choulet, F., Roger, D., Gauthier, V., Martinant, J.P., Sourdille, P., Balfourier, F., Le Paslier, M.C., Chauveau, A., et al. (2010). Insertion site-based

polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol. J. **8:** 196-210.

Paz, M.F., Wei, S., Cigudosa, J.C., Rodriguez-Perales, S., Peinado, M.A., Huang, T.H.-M., and Esteller, M. (2003). Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. Hum. Mol. Genet. **12:** 2209-2219.

Pellicer, J., Fay, M.F., and Leitch, I.J. (2010). The largest eukaryotic genome of them all? Bot. J. Linn. Soc. **164:** 10-15.

Peng, J.H., Zadeh, H., Lazo, G.R., Gustafson, J.P., Chao, S., Anderson, O.D., Qi, L.L., Echalier, B., Gill, B.S., Dilbirligi, M., et al. (2004). Chromosome Bin Map of Expressed Sequence Tags in Homoeologous Group 1 of Hexaploid Wheat and Homoeology With Rice and Arabidopsis. Genetics **168:** 609-623.

Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol. **5:** R79.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. (2006). Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Mol. Phylogenet. Evol. **39:** 70-82.

Plant DNA C-values database. (2010). http://data.kew.org/cvalues/.

Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2008). Gene expression quantitative trait locus analysis of 16000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J. **53:** 90-101.

Prokaryote Genome Size Database. (2010). http://www.genomesize.com/prokaryotes.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J. Exp. Bot. **56:** 1-14.

Qi, L.L., Friebe, B., and Gill, B.S. (2002). A strategy for enhancing recombination in proximal regions of chromosomes. Chromosome Res. **10:** 645-654.

Qi, L.L., Echalier, B., Chao, S., Lazo, G.R., Butler, G.E., Anderson, O.D., Akhunov, E.D., Dvorak, J., Linkiewicz, A.M., Ratnasiri, A., et al. (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics **168:** 701-712.

Rabinowicz, P.D., Citek, R., Budiman, M.A., Nunberg, A., Bedell, J.A., Lakey, N., O'Shaughnessy, A.L., Nascimento, L.U., McCombie, W.R., and Martienssen,

R.A. (2005). Differential methylation of genes and repeats in land plants. Genome Res. **15:** 1431-1440.

Rabinowicz, P.D., and Bennetzen, J.L. (2006). The maize genome as a model for efficient sequence analysis of large plant genomes. Curr. Opin. Plant Biol. **9:** 149-156.

Randhawa, H.S., Dilbirligi, M., Sidhu, D., Erayman, M., Sandhu, D., Bondareva, S., Chao, S., Lazo, G.R., Anderson, O.D., Miftahudin, et al. (2004). Deletion Mapping of Homoeologous Group 6-Specific Wheat Expressed Sequence Tags. Genetics **168:** 677-686.

Ravel, C., Praud, S., Murigneux, A., Canaguier, A., Sapet, F., Samson, D., Balfourier, F., Dufour, P., Chalhoub, B., Brunel, D., et al. (2006). Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (Triticum aestivum L.). Genome **49:** 1131-1139.

Ravel, C., Praud, S., Canaguier, A., Dufour, P., Giancola, S., Balfourier, F., Chalhoub, B., Brunel, D., Linossier, L., Dardevet, M., et al. (2007). DNA sequence polymorphisms and their application to bread wheat quality. Euphytica **158:** 331-336.

Ren, X.Y., Fiers, M.W., Stiekema, W.J., and Nap, J.P. (2005). Local coexpression domains of two to four genes in the genome of Arabidopsis. Plant Physiol. **138:** 923-934.

Ren, X.Y., Stiekema, W.J., and Nap, J.P. (2007). Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. Plant Mol. Biol. **65:** 205-217.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64-69.

Rizzon, C., Ponger, L., and Gaut, B.S. (2006). Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in *Arabidopsis* and Rice. PLoS Comput. Biol. **2:** e115.

Rustenholz, C., Hedley, P.E., Morris, J., Choulet, F., Feuillet, C., Waugh, R., and Paux, E. (2010). Article n°2. BMC Genomics.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16:** 944-945.

Sabot, F., Simon, D., and Bernard, M. (2004). Plant transposable elements, with an emphasis on grass species. Euphytica **139:** 227-247.

Sabot, F., Guyot, R., Wicker, T., Chantret, N., Laubin, B., Chalhoub, B., Leroy, P., Sourdille, P., and Bernard, M. (2005). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. Mol. Genet. Genomics **274:** 119-130.

Sabot, F., and Schulman, A.H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity **97:** 381-388.

Safar, J., Bartos, J., Janda, J., Bellec, A., Kubalakova, M., Valarik, M., Pateyron, S., Weiserova, J., Tuskova, R., Cihalikova, J., et al. (2004). Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J. **39:** 960-968.

Saintenac, C., Falque, M., Martin, O.C., Paux, E., Feuillet, C., and Sourdille, P. (2009). Detailed Recombination Studies along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (Triticum aestivum L.). Genetics **181:** 393-403.

Salina, E.A., Sergeeva, E.M., Adonina, I.G., Shcherban, A.B., Afonnikov, D.A., Belcram, H., Huneau, C., and Chalhoub, B. (2009). Isolation and sequence analysis of the wheat B genome subtelomeric DNA. BMC Genomics **10:** 414.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (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., Chague, V., Bolot, S., Magdelenat, G., Huneau, C., Pont, C., Belcram, H., Couloux, A., Gardais, S., Evrard, A., et al. (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.

Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T.J., Messing, J., and Feuillet, C. (2009). Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc. Natl. Acad. Sci. U. S. A. **106:** 14908-14913.

Sandhu, D., Champoux, J.A., Bondareva, S.N., and Gill, K.S. (2001). Identification and physical localization of useful genes and markers to a major gene-rich region on wheat group 1S chromosomes. Genetics **157:** 1735-1747.

Sandhu, D., and Gill, K.S. (2002). Gene-Containing Regions of Wheat and the Other Grass Genomes. Plant Physiol. **128:** 803-811.

Sanmiguel, P., and Bennetzen, J.L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. **82:** 37-44.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43-45.

Santos, A.P., Abranches, R., Stoger, E., Beven, A., Viegas, W., and Shaw, P.J. (2002). The architecture of interphase chromosomes and gene positioning are altered by changes in DNA methylation and histone acetylation. J. Cell Sci. **115:** 4597-4605.

Sato, K., Nankaku, N., and Takeda, K. (2009). A high-density transcript linkage map of barley derived from a single population. Heredity **103:** 110-117.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of Arabidopsis thaliana development. Nat. Genet. **37:** 501-506.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463:** 178-183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112-1115.

Sears, E.R. (1954). The aneuploids of common wheat. Missouri Agricultural Experiment Station. Research Bulletin **572:** 1-58.

Sears, E.R., and Sears, L. (1978). The telocentric chromosomes of common wheat In Proc. 5th Int. Wheat Genetics Symp., S. Ramanujams, ed (New Delhi, India. : Indian Agricultural Research Institute), pp. 389-407.

See, D.R., Brooks, S., Nelson, J.C., Brown-Guedira, G., Friebe, B., and Gill, B.S. (2006). Gene evolution at the ends of wheat chromosomes. Proc. Natl. Acad. Sci. U. S. A. **103:** 4162-4167.

Sémon, M., and Duret, L. (2006). Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals. Mol. Biol. Evol. **23:** 1715-1723.

Shaw, P.J. (2010). Mapping chromatin conformation. F1000 Biol. Rep. **2:** 18.

Shoja, V., and Zhang, L. (2006). A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat. Mol. Biol. Evol. **23:** 2134-2141.

Sidhu, D., and Gill, K.S. (2004). Distribution of genes and recombination in wheat and other eukaryotes. Plant Cell Tiss. Org. **79:** 257-270.

Soderlund, C., Longden, I., and Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. **13:** 523-535.

Soderlund, C., Humphray, S., Dunham, A., and French, L. (2000). Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. **10:** 1772-1787.

Somers, D.J., Kirkpatrick, R., Moniwa, M., and Walsh, A. (2003). Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome **46:** 431-437.

Spellman, P.T., and Rubin, G.M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. **1:** 5.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. **5:** e1000734.

Stein, N., Prasad, M., Scholz, U., Thiel, T., Zhang, H.N., Wolf, M., Kota, R., Varshney, R.K., Perovic, D., Grosse, I., et al. (2007a). A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. Theor. Appl. Genet. **114:** 823-839.

Stein, N., Prasad, M., Scholz, U., Thiel, T., Zhang, H.N., Wolf, M., Kota, R., Varshney, R.K., Perovic, D., Grosse, I., et al. (2007b). A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. Theor Appl Genet **114:** 823-839.

Stolc, V., Li, L., Wang, X., Li, X., Su, N., Tongprasit, W., Han, B., Xue, Y., Li, J., Snyder, M., et al. (2005). A pilot study of transcription unit analysis in rice using oligonucleotide tiling-path microarray. Plant Mol. Biol. **59:** 137-149.

Suchánková, P., Kubaláková, M., Kovářová, P., Bartoš, J., Číhalíková, J., Molnár-Láng, M., Endo, T., and Doležel, J. (2006). Dissection of the nuclear genome of barley by chromosome flow sorting. Theor Appl Genet **113:** 651-659.

Sweredoski, M., DeRose-Wilson, L., and Gaut, B.S. (2008). A comparative computational analysis of nonautonomous helitron elements between maize and rice. BMC Genomics **9:** 467.

Swift, H.H. (1950). The constancy of desoxyribose nucleic acid in plant nuclei. Proc. Natl. Acad. Sci. U. S. A. **36:** 643-654.

Szekvolgyi, L., and Nicolas, A. (2010). From meiosis to postmeiotic events: homologous recombination is obligatory but flexible. FEBS J. **277:** 571-589.

Tang, H., Bowers, J.E., Wang, X., and Paterson, A.H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. U. S. A. **107:** 472-477.

Tenaillon, M.I., Hollister, J.D., and Gaut, B.S. (2010). A triptych of the evolution of plant transposable elements. Trends Plant Sci. **15:** 471-478.

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science **327:** 818-822.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-814.

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

Thiel, T., Graner, A., Waugh, R., Grosse, I., Close, T.J., and Stein, N. (2009). Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. BMC Evol. Biol. **9:** 209.

Thimmapuram, J., Duan, H., Liu, L., and Schuler, M.A. (2005). Bicistronic and fused monocistronic transcripts are derived from adjacent loci in the Arabidopsis genome. RNA **11:** 128-138.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S., and Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. **19:** 2221-2230.

Tirichine, L., Andrey, P., Biot, E., Maurin, Y., and Gaudin, V. (2009). 3D fluorescent in situ hybridization using Arabidopsis leaf cryosections and isolated nuclei. Plant Methods **5:** 11.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. Nature **461:** 423-426.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313:** 1596-1604.

Varshney, R.K., Sigmund, R., Börner, A., Korzun, V., Stein, N., Sorrells, M.E., Langridge, P., and Graner, A. (2005). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Science **168:** 195-202.

Varshney, R.K., Hoisington, D.A., and Tyagi, A.K. (2006). Advances in cereal genomics and applications in crop breeding. Trends Biotechnol. **24:** 490-499.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., et al. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). Nat. Genet. **42:** 833-839.

Vitte, C., and Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet. Genome Res. **110:** 91-107.

Vitte, C., and Bennetzen, J.L. (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. U. S. A. **103:** 17638-17643.

Vitte, C., Panaud, O., and Quesneville, H. (2007). LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics **8**.

Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., Xiao, J., et al. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. Plant J. **61:** 752-766.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. **37:** 853-862.

Wei, F., Stein, J.C., Liang, C., Zhang, J., Fulton, R.S., Baucom, R.S., De Paoli, E., Zhou, S., Yang, L., Han, Y., et al. (2009). Detailed analysis of a contiguous 22-Mb region of the maize genome. PLoS Genet. **5:** e1000728.

Wicker, T., Matthews, D.E., and Keller, B. (2002). TREP: a database for Triticeae repetitive elements. Trends Plant Sci. **7:** 561-562.

Wicker, T., Zimmermann, W., Perovic, D., Paterson, A.H., Ganal, M., Graner, A., and Stein, N. (2005). A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats. Plant J. **41:** 184-194.

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. **17:** 1072-1081.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8:** 973-982.

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T., Graner, A., Ware, D., and Stein, N. (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics **9:** 518.

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M., and Stein, N. (2009). A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J. **59:** 712-722.

Wicker, T., Buchmann, J.P., and Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. **20:** 1229-1237.

Williams, E.J., and Bowles, D.J. (2004). Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res. **14:** 1060-1067.

Williams, P.C. (1993). The world of wheat. In Grains and Oilseeds: Handling, Marketing, Processing (Canadian International Grains Institute, Winnipeg, Manitoba, Canada).

Wright, S.I., Agrawal, N., and Bureau, T.E. (2003). Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. Genome Res. **13:** 1897-1903.

Xu, Z., Kohel, R.J., Song, G., Cho, J., Alabady, M., Yu, J., Koo, P., Chu, J., Yu, S., Wilkins, T.A., et al. (2008). Gene-rich islands for fiber development in the cotton genome. Genomics **92:** 173-183.

Zhan, S., Horrocks, J., and Lukens, L.N. (2006). Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. Plant J. **45:** 347-357.

Zhang, L., and Gaut, B.S. (2003). Does Recombination Shape the Distribution and Evolution of Tandemly Arrayed Genes (TAGs) in the Arabidopsis thaliana Genome? Genome Res. **13:** 2533-2540.

Zilberman, D., and Henikoff, S. (2007). Genome-wide analysis of DNA methylation patterns. Development **134:** 3959-3965.

Zonneveld, B.J., Leitch, I.J., and Bennett, M.D. (2005). First nuclear DNA amounts in more than 300 angiosperms. Ann. Bot. **96:** 229-244.