Master II Biologie et Environnement

Génomique, Ecophysiologie et Productions Végétales

Présenté par
Josquin Daron

Caractérisation de gènes de blé non synténiques avec les génomes des espèces apparentées

Responsables du stage :
Frédéric Choulet (Ingénieur de Recherche)
Catherine Feuillet (Directrice de Recherche)
Juin 2011

Sommaire

Liste des abréviations

ABI Solid : Sequencing by Oligonucleotide Ligation and Detection

ADNc : ADN complémentaire

BAC : Bacterial Artificial Chromosomes

BLAST : Basic Local Alignment Search Tool (protéine

CDS : Coding DNA Sequence

CS5x : Couverture de 5x du génome de Chinese Spring

EST : Expressed Sequence Tag

ET : Eléments transposables

Gb : gigabases

kb : kilobase

LTR : Long Terminal Repeat sequence

Mb :mégabase

ORF : Open reading fram

PCR :Polymerase chaine réaction

SNP : Single-Nucleotide Polymorphism

Tm : melting temperature (température de fusion)

UTR : UnTranslated Region

WGS : Whole Genome Shotgun

PREAMBULE

Depuis plusieurs années, la génomique révolutionne nos connaissances sur le développement et le métabolisme des plantes. Avec le séquençage des génomes et le génotypage à haut débit, de nouvelles perspectives s'ouvrent en production variétale, notamment chez la famille des Poaceae (graminées) qui regroupe des espèces cultivées d'importance majeure. Toutefois, contrairement aux espèces comme le riz, le sorgho ou même le maïs, les approches génomiques chez les *Triticeae* (blé, orge et seigle) s'avèrent plus difficiles, à cause de leur génome composé entre 80 et 90% de séquences d'ADN répétées, dérivées d'éléments transposables (Flavell et al. 1974) et d'une taille très élevée (plusieurs Gb). Le génome du blé tendre est allohexaploïde, c'est-à-dire qu'il est composé de trois génomes homéologues A, B et D (avec 2n=6x=AABBDD) dont la taille totale est estimée à 17 Gb, soit 40 fois plus grand que celui du riz et plus de 100 fois plus grand que celui d'Arabidopsis.

En outre, l'un des défis majeurs de la génomique actuelle est d'établir une séquence de référence pour le génome du blé. En effet, ce génome reste, encore aujourd'hui, très mal caractérisé et diverses hypothèses, parfois contradictoires, ont été émises quant à sa composition et notamment, concernant la distribution des gènes le long des chromosomes.

Ainsi, un consortium international a été mis en place en 2005 : l'IWGSC (www.wheatgenome.org) Consortium International pour le Séquençage du Génome du Blé, au sein duquel la stratégie adoptée vise à réduire la complexité génomique, en triant les chromosomes par cytométrie en flux. Le chromosome 3B représentant 1000 Mb, soit à lui seul plus de 2 génomes de riz, a ainsi été utilisé afin de démontrer la faisabilité de l'approche chromosome-spécifique. En 2008, l'Unité GDEC a publié la première carte physique d'un chromosome de blé, le 3B (Paux et al. 2008), ouvrant ainsi la voie à l'établissement des cartes physiques des 20 autres chromosomes par les différents acteurs du Consortium. Grâce à cette première carte, à l'époque constituée de 56000 BAC assemblés et ordonnés en 1036 contigs, il a été possible de produire les premières grandes séquences contiguës du génome de blé (Choulet et al. 2010). Cette étude a certes, permis de mieux caractériser l'espace génique et l'organisation des éléments transposables, mais a surtout posé de nouveaux questionnements concernant l'évolution du génome. En effet, près de la moitié du contenu en gènes était inattendue sur la base de la synténie avec les génomes apparentés. L'objectif de ce stage a donc été de caractériser ces gènes issus de réarrangements récents.

Espèces	Taille du génomes (Mb)	Nombre de chromosomes	Nombre de gènes	Composition en retrotransposons (%)	Composition en transposons à ADN (%)
Brachypodium	320	2n = 2x =10	25 532	23,3	4,8
Riz	400	2n = 2x =24	28 236	25,8	13,7
Sorgho	800	2n = 2x =20	27 640	54,5	7,5
Maïs	2500	2n = 2x =20	32 540	75,9	8,6

Table 1 : Caractéristiques structurelles des quatre génomes des graminées séquencés (d'après Devos 2010).

SYNTHESE BIBLIOGRAPHIQUE

1. Composition et organisation du génome de Poaceae

a) Taille des génomes et polyploïdie

Les génomes de plantes se caractérisent par une très grande variabilité de taille et d'organisation des séquences. Par exemple, l'algue *Ostreococcus tauri* possède un génome très compact de 10 Mb, taille similaire à celle des génomes bactériens, alors que le génome de la monocotylédone *Paris Japonica* atteint les 150000 Mb (Bennett and Leitch 2011). Au sein du règne végétal, la division des Magnoliophyta, couramment appelée angiospermes ou "plantes à fleurs", présente la variabilité la plus considérable, avec une amplitude d'un facteur 2000 entre *Genlisea margaretae* (60 Mb) et *Fritilaria assyriaca* (120.000 Mb) (Bennett and Leitch 2011).

Parmi les *Magnoliophyta*, les *Poaceae* (appelées graminées) (>10000 espèces identifiées) représentent la famille la plus importante pour l'Homme des plantes cultivées. De par leur adaptabilité à des environnements très différents, les graminées constituent une famille à très forte diversité génomique qui touche aussi bien le nombre de chromosomes (allant de 2n = 4 pour *Zingeria biebersteiniana* à 2n = 266 pour le polyploïde *Poa litorosa*), la taille (allant de 390 Mb pour le riz à 17 Gb pour le blé), le niveau de ploïdie et la composition en gènes et éléments transposables) (pour revue (Keller and Feuillet 2000)).

b) Composition en gènes et éléments transposables

Les génomes de quatre espèces de Poaceae ont été complètement séquencés et sont publiquement disponibles : riz (390 Mb) (2005), sorgho (700 Mb) (Paterson et al. 2009), maïs (2500 Mb) (Schnable et al. 2009) et brachypodium (280 Mb) (2010). Leur analyse a permis d'évaluer la composition et l'organisation des gènes et des éléments répétés qui diffèrent dans les génomes compacts (<500 Mb) et dans les génomes plus grands.

Tout d'abord, le nombre de gènes codant des protéines semble être relativement conservé quelque soit la taille du génome : 25532 pour Brachypodium, 28236 chez le riz, 27640 chez le sorgho et, enfin, 32540 chez le maïs (Table 1). Cette estimation est légèrement plus élevée car le génome du maïs a subi récemment un événement de tetraploïdisation. Le nombre de familles de gènes est similaire entre ces différentes espèces. De plus, certaines familles de gènes, comme par exemple les familles synthétisant la cellulose, les protéines de stockage et les gènes de résistance aux maladies, se sont développées et diversifiées au sein des lignées des Poaceae (Devos 2010).

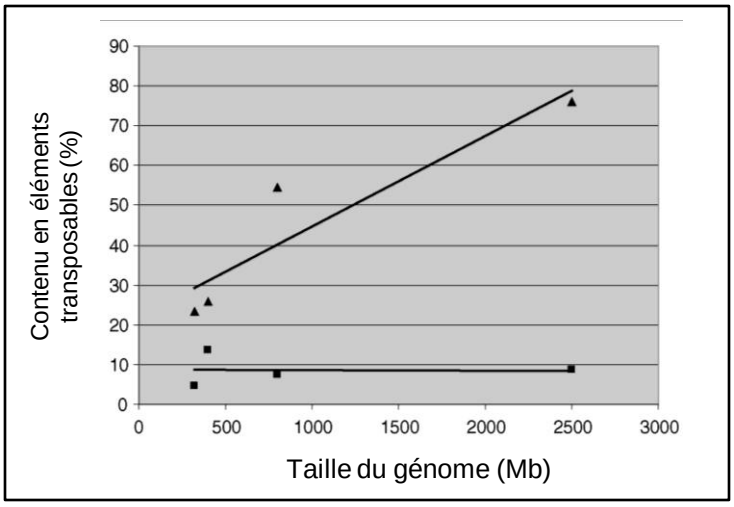


Figure 1 : Proportion de rétrotransposons à LTR (classe 1, triangles) et transposons à ADN (class 2, carrés) en fonction de la taille des génomes de B*rachypodium*, riz, sorgho et maïs (d'après Devos 2010).

Mais comme pour la majorité des plantes, la majeure partie du génome des graminées est constituée d'éléments transposables (ET) ou mobilisables capables de se multiplier au sein du génome (Feschotte and Pritham 2009). Bien que les familles d'ET retrouvées dans les génomes soient spécifiques de chaque espèce (Vitte and Bennetzen 2006), la classe des retrotransposons à LTR (Long Terminal Repeats, classe I), c'est-à-dire les éléments apparentés aux rétrovirus, apparaît comme la plus représentée chez l'ensemble des graminées (Devos 2010). À l'inverse, chez les mammifères, les retrotransposons sans LTR (LINE et SINE) sont les plus massivement représentés (Gregory et al. 2007). La proportion de retrotransposons à LTR est positivement corrélée à la taille des génomes, démontrant qu'ils sont les principaux responsables de l'expansion des génomes (Figure 1). Ils représentent 23% du génome de Brachypodium, 26% de celui du riz, 55% chez le sorgho, 76% du génome de maïs et leur proportion est estimée aux alentours de 70% chez les génomes du blé, de l'orge et du seigle. L'expansion des génomes est donc dépendante de l'activité des éléments transposables. Chez le riz par exemple, un petit nombre de familles de retrotransposons à LTR ont, à elles seules, et en moins de 3 millions d'années, causé l'expansion massive du génome de l'espèce sauvage *Oryza australiensis* : 965 Mb, comparé à 350-500 Mb pour les espèces voisines (Piegu et al. 2006). Chez les *Triticeae*, de la même manière, deux familles de Gypsy (Fatima et Sabrina) seulement sont responsables de l'expansion du génome B (Paux et al. 2006).

Par ailleurs, les transposons à ADN de type MITE (classe II) sont également retrouvés en un grand nombre de copies dans l'ensemble des génomes de graminées séquencés (Devos 2010). Au vu de la grande variation de taille et de composition en ET des génomes des graminées, la distribution des gènes le long des chromosomes suit deux tendances : pour les génomes compacts tels que celui du riz, du sorgho et de brachypodium, les régions géniques sont réparties de façon homogène dans l'euchromatine, tandis que les ET sont concentrés dans les régions péri-centromériques (Baucom et al. 2009; Devos 2010). En revanche pour les génomes ayant subi une expansion massive (maïs, blé, orge et seigle), les gènes apparaissent comme répartis en îlots séparés par de grands clusters de séquences répétées (Liu et al. 2007). La variabilité structurale des génomes suggère une grande plasticité génomique corrélée à une forte activité de transposition, particulièrement intense pour les grands génomes de plantes.

2. Colinéarité entre les génomes des Poaceae

Les gènes des espèces relativement proches sont généralement trouvés dans un ordre similaire sur les chromosomes. Cette "colinéarité" reflète la descendance des génomes d'un ancêtre commun.

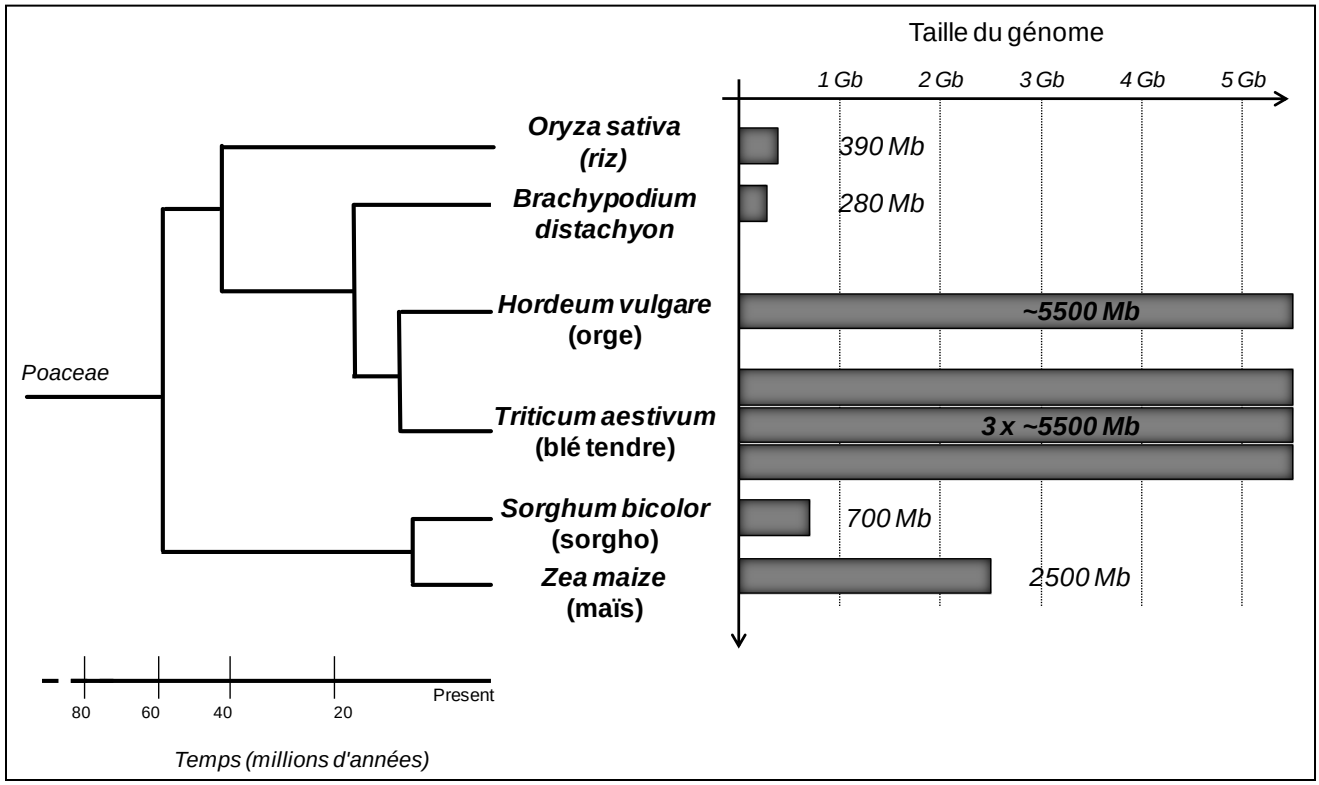


Figure 2 : Arbre phylogénétique des *Poaceae* et taille des génomes

La famille des graminées a divergé à partir d'un ancêtre commun, il y a 50 à 70 millions d'années (Prasad et al. 2005) (Figure 2). Les études réalisées dans le domaine de la paléo-génomique (reconstruction des génomes ancestraux) ont permis de mettre en évidence 9138 protogènes (les gènes présents dans le génome ancêtre) (Salse et al. 2009) et de reconstruire un génome ancêtre possédant 5 chromosomes dont dérive l'ensemble des espèces actuelles (Bolot et al. 2009). Les premières études préliminaires de génomique comparative ont montré qu'à l'échelle d'une carte génétique il existe une forte conservation de la position et de l'ordre des marqueurs, permettant de déterminer quels sont les chromosomes ou segments partagés entre les espèces (Gale and Devos 1998). Ainsi les chromosomes ou segments de chromosomes correspondants sont appelés "synténiques". Chez les poaceae, il apparaît donc que le contenu en gènes est relativement bien conservé, puisque 70% à 80% des gènes sont retrouvés conservés entre les différents génomes (Abrouk et al. 2010). Parallèlement à cette forte conservation, la structure des gènes est elle aussi fortement conservée : les récentes comparaisons d'EST ("expressed sequence tags") du riz, du blé, de l'orge, de sorgho, et du maïs ont montré que ~70% des gènes maintiennent une structure commune (Bolot et al. 2009).

En revanche, l'identification de la position des gènes conservés entre ces différentes espèces permet de constater que seulement 16% des gènes en moyenne sont conservés en position de synténie entre le riz, *Brachypodium*, le sorgho, le maïs l'orge et le blé (Abrouk et al. 2010).Plus précisément, entre les génomes du riz et du sorgho 57% des gènes ont été retrouvés en position synténique (Paterson et al. 2009). De même, l'alignement de 656 gènes orthologues entre le riz et le maïs a montré que, respectivement, 27,2% et 21,8% d'entre eux n'étaient pas retrouvés en position synténique chez le blé (Bolot et al. 2009).

En ce qui concerne le blé, les analyses de présence ou d'absence de loci orthologues au sein des trois sous génomes A, B et D, ont révélé que non seulement la synténie entre les chromosomes homéologues a été perturbée par de fréquentes insertions et délétions, mais, qu'en plus, cette diminution est corrélé à la distance au centromère (Akhunov et al. 2003). Ainsi, environ 90% du génome B pour seulement 95% pour les génomes A et D, sont apparus en relation de synténie.

Les régions non-codantes et inter-géniques des génomes apparaissent fortement divergentes dans la famille des graminées. En effet, des insertions ou délétion d'ET causés par des mécanismes de recombinaison inégale ou illégitime, remanient complètement ces régions en quelques millions d'années (Devos et al. 2002), (SanMiguel et al. 2002), (Wicker et al. 2003).

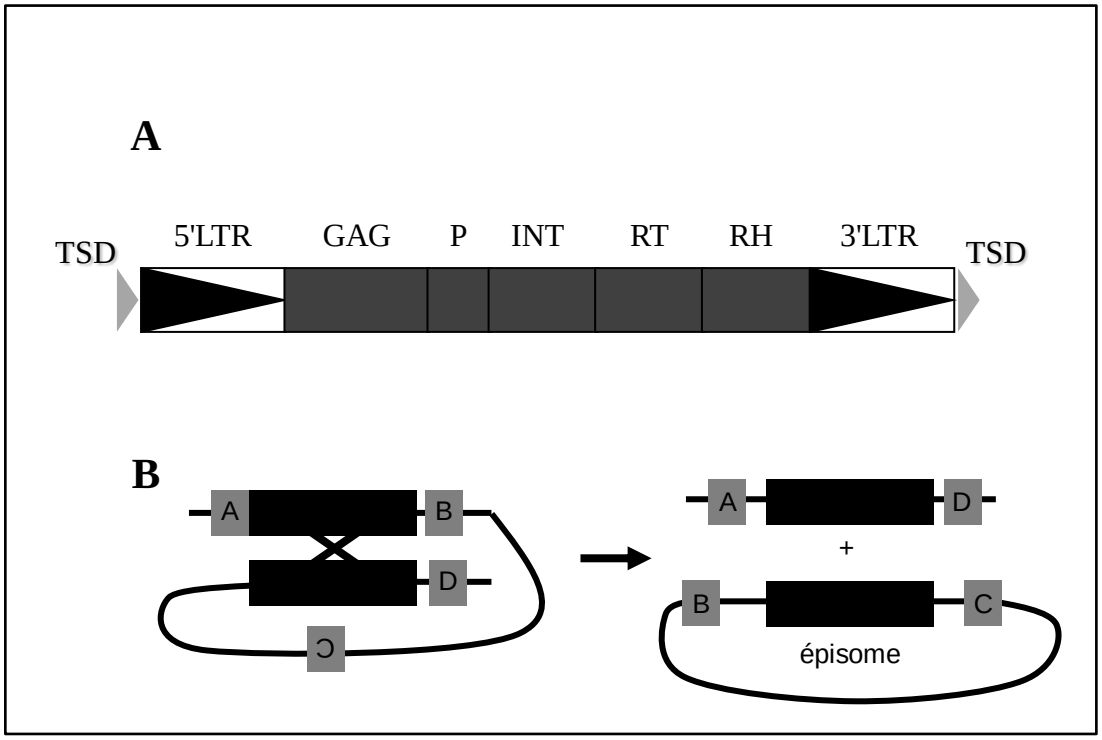


Figure 3 : Structure des rétrotransposons à LTR et implication dans la plasticité génomique.
A. Structure d'un rétrotransposon à LTR (LTR : "Long Terminal Repeats", GAG : protéine de liaison à l'ADN, P : Protéase, INT : Intégrase, RT : Réverse Transcriptase, RH : RNAse H, TSD : "target site duplication"). B. Exemple de recombinaison illégitime entre 2 éléments transposables (en noir) . L'événement génère un épisome porteur des gènes B et C qui sera perdu au cours des générations suivantes.

En effet, libre de toute pression de sélection, les régions non-codantes évoluent rapidement (Petrov 2001).

A l'échelle des chromosomes, il apparaît que la colinéarité au sein des Poaceae a fortement été conservée au cours des 60 millions d'années d'évolution. Mais les nombreux réarrangements qui opèrent au sein du génome des plantes ne peuvent-ils pas, au cours des âges, éroder cette cohérence génomique, comme ce qui est observé entre les monocotylédones et les dicotylédones (Paterson et al. 2004)?

3. Mécanismes moléculaires et forces évolutives impliqués dans les réarrangements génomiques

a) Mécanisme de réarrangement génomique

Il est possible de distinguer quatre mécanismes moléculaires courants à l'origine des réarrangements génomiques chez les plantes : la polyploïdisation, la transposition d'éléments mobiles, la recombinaison illégitime (Figure 3) et la recombinaison inégale homologue (Bennetzen 2007). Ces mécanismes sont responsables de l'expansion (amplifications/duplications) et de la contraction (délétions) des génomes. Par exemple, la polyploïdisation est un événement de duplication de l'ensemble du génome créant une redondance en gènes au sein d'un même génome (Hufton and Panopoulou 2009).

b) Les duplications des gènes

Le mouvement des gènes a conduit à des bouleversements de la colinéarité des génomes mettant en lumière des mécanismes de micro réarrangement, aboutissant à la création de gènes "orphelins" et/ou de gènes "nouveaux".

Mobilisation de gènes et "shuffling" d'exons

Les ET sont souvent considérés comme les principaux responsables des réarrangements des gènes (Kapitonov and Jurka 2007). En effet, il apparaît que certains ET peuvent lors de leur transposition, acquérir des séquences génétiques spécifiques et les amplifier sur le génome (Kapitonov and Jurka 2007). Le mécanisme de la capture de gènes ou de fragments de gène le long du génome a été appelé mécanisme d'exon "shuffling" (Wicker et al. 2010). Ce phénomène a été décrit pour les éléments de la superfamille des Mutator (Jiang et al. 2004), Helitron (Lai et al. 2005), CACTA (Wicker et al. 2003), (Paterson et al. 2009) et LTR retrotransposons (Bennetzen 2007).

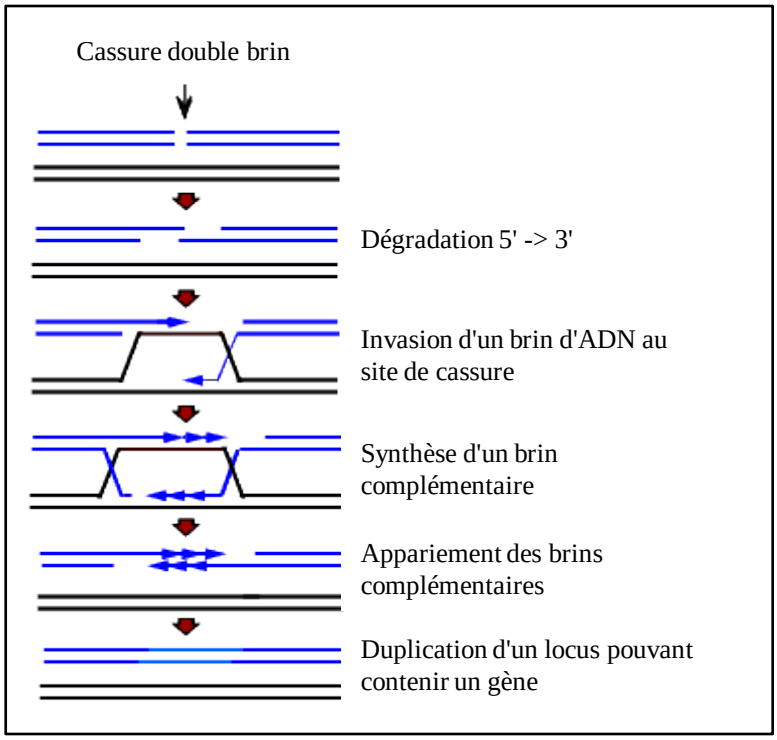


Figure 4 : Réparation d'une cassure double brin par SDSA ("Synthesis-Dependant Strand Annealing")

Un événement de cassure de l'ADN double brin survient (après l'insertion d'un élément transposable par exemple) dans le génome. Un brin d'ADN étranger pouvant porté un gène est utilisé comme matrice pour la synthèse d'un brin complémentaire, aboutissant à la duplication d'un locus.

Les premiers mécanismes de capture d'ADN ont été observés pour les ET Mutator (Mutator like DNA Element, MULEs) chez Arabidopsis et le maïs (Yu et al. 2000). Cette famille d'ET apparaît comme largement impliquée dans réarrangement d'ADN, puisque 3000 MULEs contenant des fragments d'ADN entre 47-986 bp ont été découverts et appelés "Pack-MULEs" (Jiang et al. 2004). Au moins, 20% des fragments des gènes contenus par les Pack-MULEs ont été observés comme transcrits. De plus 1% de ces fragments a été analysé comme "nouveau", suggérant le rôle des Pack-MULEs dans la création de nouveaux gènes (Hanada et al. 2009). En outre, il semblerait que ces éléments représentant 1,6% du génome du riz, soient largement représentés parmi les génomes des angiospermes.

Une autre famille d'ET identifiée comme majoritairement impliquée dans le transport de gènes est celle des *Helitron*. Pour cause de détection difficile, cette famille n'a que très récemment été identifiée, et n'a été caractérisée que chez les génomes de plantes. Ainsi, il apparaît chez le maïs que les *Helitron* occuperaient 2,2% du génome, avec 60% des éléments complets contenant un gène ou plus (Yang and Bennetzen 2009). Ainsi la capture de gènes par les ET est de plus en plus considérée comme une stratégie de contournement des mécanismes de répression mise en place par l'organisme hôte contre les ET (Tenaillon et al. 2010).

Réparation de cassures d'ADN double-brin par SDSA : "Synthesis-Dependent Strand Annealing")

La capture de gènes par les ET n'est pas l'unique mécanisme de la capture de gènes. Wiker et al., ont comparés les génomes du riz, de *Brachypodium* et du sorgho afin d'identifier des gènes non colinéaires issus de duplications (Wicker et al. 2010). Cette étude a montré la présence de grandes régions dupliquées (jusqu'à 50kb) bornant un ET et portant des gènes. Cela suggère que les ET ne seraient pas les responsables directs de ces captures de fragments de gènes, mais ils seraient les initiateurs de cassures double brin lors de leur insertion. L'hypothèse émise est que ces duplications seraient causées par la réparation de telles cassures. Lors de réparations imparfaites, le fragment d'ADN utilisé pour combler la cassure pourrait porter un gène ou un fragment de gène (Figure 4) (Wicker et al. 2010) un mécanisme nommé "Synthesis-Dependent Strand Annealing".

Il apparaît que la taille des fragments d'ADN capturés par des ET serait inférieure à 1kb, alors que celle des fragments d'ADN retrouvés après une cassure double brin serait fréquemment de plusieurs kilobases (Wicker et al. 2010). Ceci expliquerait que la réparation des cassures de l'ADN double brin jouerait un rôle plus important dans les duplications de gènes. De plus, la

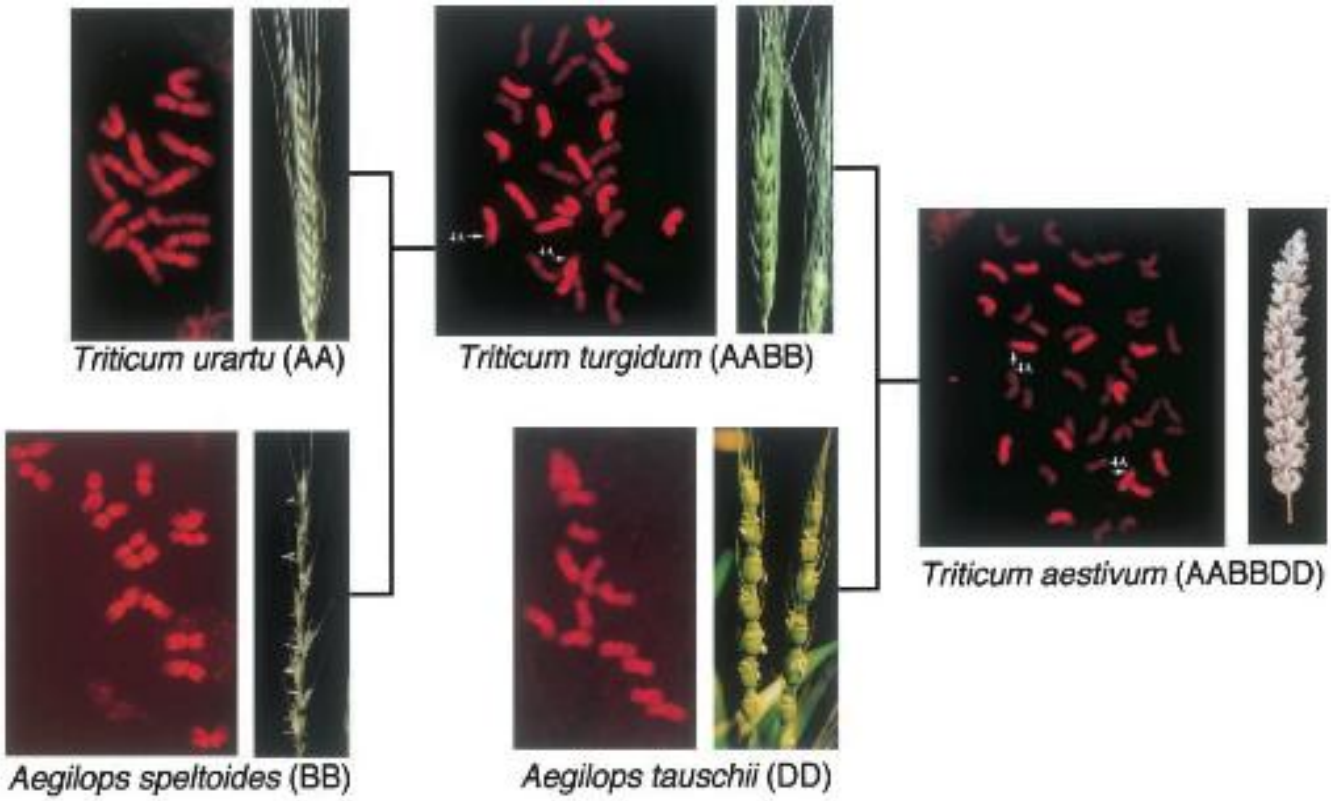


Figure 5 : Deux événements de polyploïdie à l'origine de la formation du blé tendre.
(d'après Gill et al., 2004)

fréquence plus élevée des duplications de gènes chez le riz que chez *Brachypodium* suggère que le mécanisme joue un rôle dans l'expansion des génomes

4. Structure et évolution du génome de blé

a) Le génome du blé est issu de deux événements récents de polyploïdisation

Le blé tendre (*Triticum aestivum* L.) est une espèce de la famille des *Poaceae* dont le génome est allo-hexaploïde (2n=6x=42 chromosomes) issu de la fusion de trois génomes A, B et D au sein du même noyau (Figure 5). Le blé tendre est un polyploïde très récent dans l'histoire évolutive. Cette particularité fait du blé un modèle idéal d'étude du devenir des gènes dupliqués suite à un doublement chromosomique. En effet, la perte massive de gènes dupliqués, observée généralement après un événement de duplication du génome chez les anciens polyploïdes, n'a pas encore érodé la synténie entre les trois sous-génomes. Les études comparatives démontrent que les gènes présents de façon ancestrale sont présents en trois copies chez l'hexaploïde. Ces copies de gènes dupliqués par polyploïdisation sont appelées des "gènes homéologues". Du fait de l'absence de séquence complète du génome du blé, de nombreuses études ont tenté d'estimer le nombre de gènes ainsi que leur organisation. Au vu de la forte conservation des gènes entre les génomes des céréales (Keller and Feuillet 2000), (Bolot et al. 2009)les données de séquençage de génome des espèces modèles (riz, *Brachypodium)* ont permis, dans un premier temps, d'estimer que les sous-génomes du blé devaient compter environ le même nombre de gènes soit entre 25000 et 50000 gènes (Devos et al. 2002). Quant à l'organisation de l'espace génique, si dans un premier temps a avait été perçu par l'analyse de 7000 EST cartographiés le long des chromosomes du blé (Qi et al. 2004) comme hétérogène, des études plus récente ont identifiées que cette densité en gènes suivrait un gradient positif des centromères vers les télomères des chromosomes, avec entre 15 et 20% des gènes localisés au niveau des centromères (Rustenholz et al. 2010).

La taille des génomes diploïdes de *Triticum* et d'*Aegilops* a été estimée entre 5 et 6 Gb selon les espèces. Ainsi, le génome du blé hexaploïde représente 17 Gb. Cette taille, associée à un contenu en éléments répétés supérieur à 80%, font de la génomique du blé un véritable défi pour les analyses moléculaires. Ainsi, dans le cadre du Consortium international (IWGSC), une approche visant à réduire cette complexité a été entreprise via le tri des chromosomes.

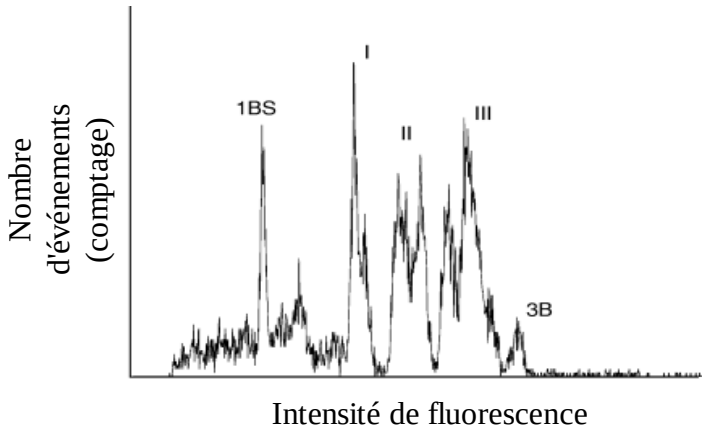


Figure 6 : Tri de chromosomes par cytométrie en flux d'un caryotype de blé tendre (lignée aneuploïde ayant perdu le bras 1BL).

Le pic "3B" contient uniquement des chromosomes 3B triés. Les autres pics représentent un mélange de chromosomes (d'après Dolezel et al., 2007). L'utilisation des lignées aneuploïdes permet de trier individuellement tous les bras de chromosomes (le 1BS sur cette figure).

b) Le tri de chromosomes et l'établissement d'une carte physique du chromosome 3B

La cytométrie en flux permet le tri des chromosomes en fonction de leur taille (Figure 6). Cette technique est notamment maîtrisée par l'équipe de Jaroslav Dolezel de "l'Institut of Experimental Botany" en République Tchèque. Toutefois, chez le blé hexaploïde, la plupart des chromosomes sont de taille très similaire, ce qui ne permet pas d'isoler tous les chromosomes individuellement. Seul le 3B (1 Gb à lui seul, le plus grand des 21 chromosomes) peut être isolé de façon individuelle car sa taille est suffisamment différente des autres chromosomes. Grâce à cette avancée, une banque BAC spécifique du chromosome 3B a pu être construite et l'Unité GDEC a entrepris, en 2005, l'établissement de la première carte physique d'un chromosome de blé via la comparaison des profils de restriction de 67968 clones BAC. Cette carte a été publiée en 2008 (Paux et al. 2008). Elle se compose de 56952 BAC (représentant une couverture de 10x) assemblés en 1036 contigs de BAC qui ont été ancrés le long du chromosome à l'aide de 1443 marqueurs moléculaires. La démonstration de faisabilité de cette approche et l'utilité de la carte physique dans les projets de clonage positionnel de gènes d'intérêts agronomiques ont permis d'initier le même travail sur tous les bras de chromosomes dans le cadre d'un effort international concerté (voir www.wheatgenome.org/Projects).

c) L'analyse des premières grandes séquences disponibles du génome de blé ont révélé la présence de nombreux gènes non colinéaires avec les espèces proches

L'établissement de la carte physique du 3B a permis de séquencer les premières grandes régions contiguës du génome de blé : 13 contigs de BAC représentant 18 Mb au total (Choulet et al. 2010). Cette étude a révélé la présence de 148 gènes codant potentiellement des protéines fonctionnelles, 27 pseudogènes et 24 fragments de gènes. Sur la base de ces données une estimation du nombre de gènes fut émise, comprise entre 6 000 et 8 400 gènes pour le chromosome 3B, et de façon plus globale, comprise entre 36 000 et 50 000 gènes pour le génome B du blé hexaploide.

De plus la répartition en gènes le long des contigs est apparue comme homogène, puisque qu'aucune région de plus de 800kb n'a été annotée sans gène, même dans la région centromérique. En revanche de fortes disparités ont été observées en fonction des contigs, car 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. À une échelle plus fine, des disparités dans la répartition des gènes ont également été constatées. Ainsi il est apparu que 75% des gènes étaient organisés en îlots.

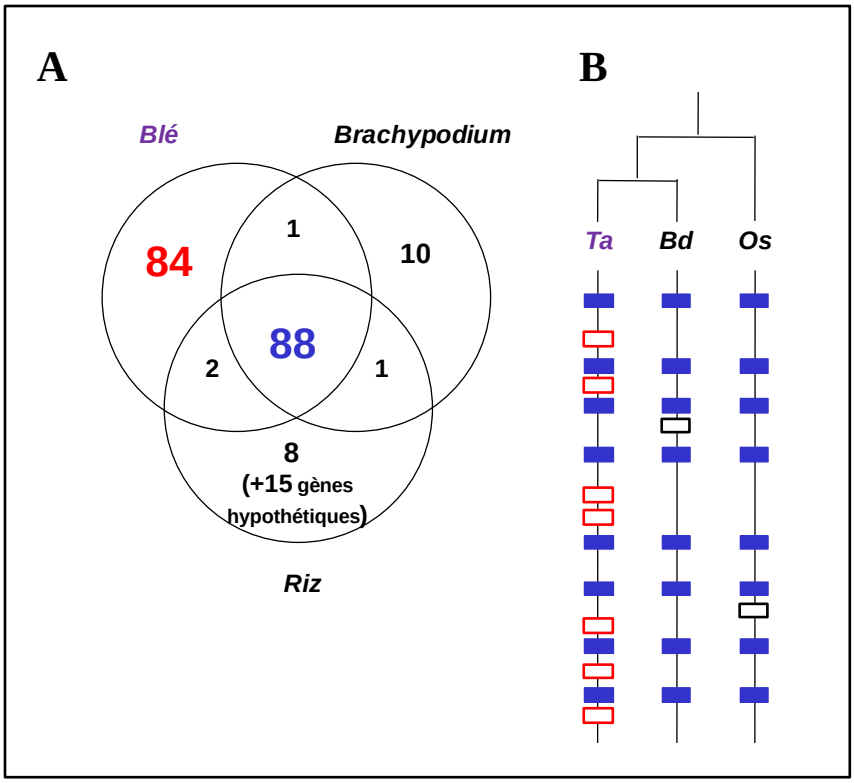


Figure 7 : Colinéarité entre les génome du blé, du riz et de *Brachypodium* identifiée après l'analyse de 13 régions du chromosome 3B.

(A) Diagramme de Venn présentant la répartition des gènes identifiés dans les 13 régions génomiques étudiées chez les génomes du blé, de Brachypodium et du riz, en fonction de leur conservation (synténiques) ou non (non synténiques) entre les 3 espèces. (B) Représentation schématique de la colinéarité entre les 3 génomes : les gènes orthologues sont représentés en bleu, les gènes non synténiques sont représentés en rouge chez le blé (Ta) et en noir chez les le riz (Os) et *Brachypodium* (Bd) (d'après Choulet et al., 2010).

Les études comparées avec les génomes modèles séquencés ont montré que 48% des gènes annotés n'ont pas d'orthologue présent sur le chromosome 1 du riz ou le chromosome 2 de *brachypodium,* c'est-à-dire en position de synténie (Figure 7). Ce chiffre, bien que nettement supérieur, est en accord avec les précédents résultats issus de la cartographie des EST, montrant que 25% des loci de gènes étaient dupliqués de manière inter chromosomique ou intra chromosomique (Akhunov et al. 2003). Ainsi ces gènes non synténiques sembleraient responsables du gradient de la densité de gènes le long du chromosome 3B.

d) L'approche du séquençage "génome entier" confirme la présence de nombreux réarrangements du contenu en gènes spécifiques des Triticeae

Depuis l'avènement des technologies de séquençage à haut débit (Roche 454, Illumina et ABI Solid), des initiatives ont été entreprises par la communauté internationale afin d'obtenir une séquence assemblée du génome via des approches "génome entier" ou "chromosome entier" ("whole shotgun" en anglais, WGS), c'est-à-dire sans passer par une étape de création d'une banque de BAC ordonnés. La colinéarité entre les génomes de *Triticeae* et les génomes modèles a pu ainsi être appréciée en comparant les séquences "shotgun" produites à partir des chromosomes triés du blé (Wicker et al. 2011) et de l'orge (Mayer et al. 2011) aux séquences des génomes complets.

Les chromosomes 1A, 1B, 1D (du blé) et 1H (de l'orge) ont été séquencés par la technologie Roche-454 à des niveaux de couverture faibles : 1 à 2x (Wicker et al 2011). Leur analyse a permis d'identifier plus de 80% des 2100 gènes attendus sur la base de la synténie avec les génomes connus. Mais de façon surprenante, près de 6000 gènes non synténiques ont aussi été observés, et près de 2000 sont conservés entre au moins 2 des génomes A, B, D et H, indiquant qu'ils ont été dupliqués avant la divergence A-B-D il y a ~3 millions d'années (Figure 8). Les gènes non synténiques représenteraient donc plus de 50% du contenu en gènes des chromosomes du groupe 1, accentuant encore les conclusions émises précédemment sur l'ampleur de la variabilité du génome des *Triticeae* (Choulet et al. 2010).

5. Objectifs du stage

Parmi les 175 gènes et pseudogènes identifiés sur les 13 contigs de BAC séquencés précédemment et portés par le chromosome 3B (Choulet et al. 2010), près de la moitié sont des gènes non-colinéaires (non-synténiques) avec les génomes du riz et de *Brachypodium*. Ceci signifie qu'ils se sont insérés à un nouveau locus (sur le 3B) récemment, c'est-à-dire après la divergence du blé et de Brachypodium, il y a moins de 20 à 30 millions d'années. L'hypothèse la

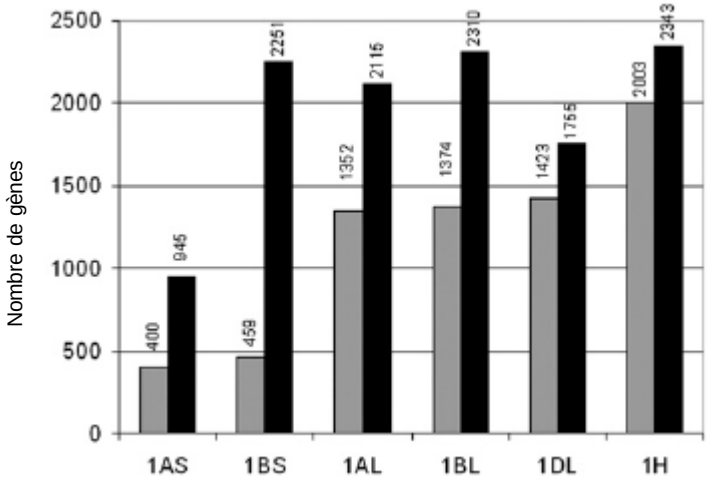


Figure 8 : Histogramme représentant le nombre de gènes identifiés dans les séquences produites pour chaque bras de chromosomes du groupe 1 du blé et pour le chromosome 1H de l'orge.

Les nombres de gènes synténiques (gris) et non synténiques (noir) avec les génomes du riz, de Brachypodium et du sorgho, identifiés par similarité avec les séquences de chaque bras de chromosomes sont indiqués pour chaque bras. La lignée aneuploïde 1DS présente un réarrangement chromosomique, et aucun résultat n'est disponible pour celle-ci (d'après Wicker et al., 2011).

plus probable de leur origine fait intervenir un événement de duplication inter-chromosomique d'un gène ancestralement présent à un locus non-synténique.

L'objectif principal de ce stage est donc de répondre aux questions suivantes : les gènes non-colinéaires sont-ils issus d'événement de duplication inter-chromosomiques ou de translocations ? Combien de copies de ces gènes sont présents dans le génome du blé ? Sont-ils exprimés ou sont-ils des pseudogènes ? Dans le but de répondre à ces différentes questions, une approche d'analyse bioinformatique de séquences a été entreprise. L'un des objectifs était donc de développer des algorithmes bioinformatiques afin d'automatiser le traitement des séquences et l'analyse des données. Enfin, les analyses réalisées avaient pour objectif de développer des amorces conservées entre les différentes copies de gènes dupliqués, afin d'identifier, par approche de biologie moléculaire, leur localisation chromosomique et de confronter les résultats obtenus au laboratoire avec les analyses de séquences.

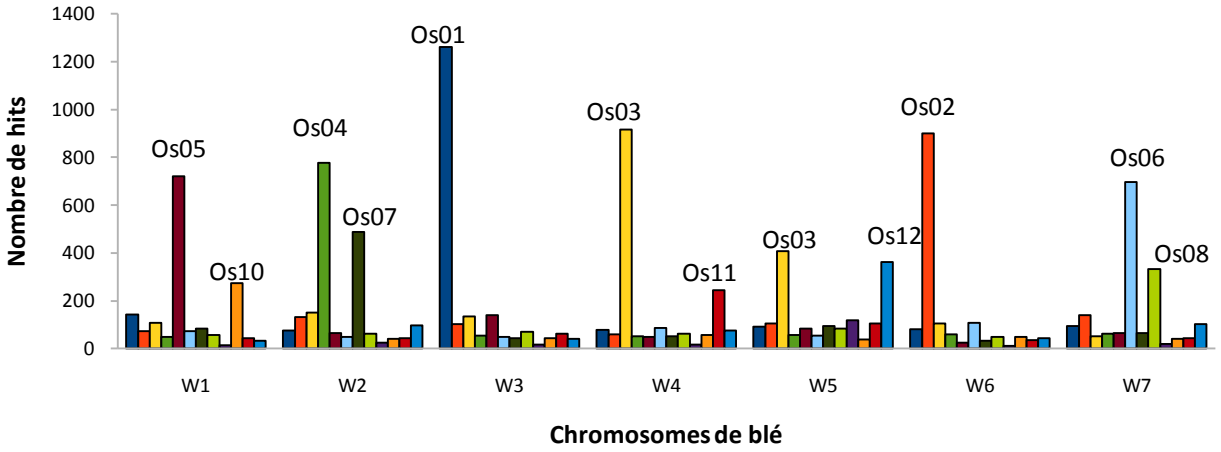


Figure 9 : Histogramme représentant la localisation chromosomique des gènes du riz homologues aux 6426 EST cartographiées sur les 7 chromosomes du blé.

Les 7 chromosomes du blé sont représentés sur l'axe des abscisses. Une recherche de similarité a été réalisée, pour 6424 EST du blé cartographiées, avec le génome du riz. Seuls les meilleurs hit de BLAST (gènes de riz) ont été retenus pour chaque EST, et une représentation des chromosomes d'origine est donnée sous forme d'histogramme. Par exemple, la majorité des EST cartographiées sur les chromosomes 3A-B-D sont similaires au chromosome 1 du riz, confirmant que ces chromosomes sont orthologues.

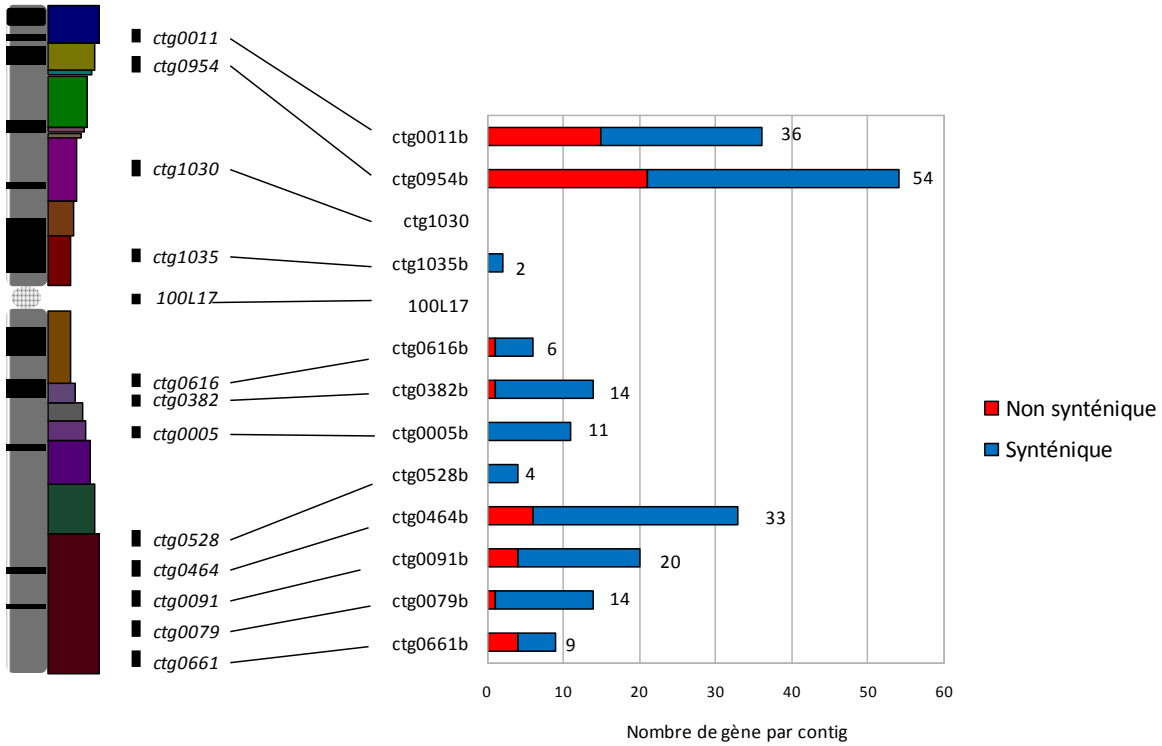


Figure 10 : Localisation des 13 contigs de BAC séquencés sur le chromosome 3B et distribution du nombre de gènes synténiques et non synténiques.

TRAVAIL EXPERIMENTAL

1. Identification des régions de synténie entre les génomes du blé, du riz et de Brachypodium

En l'absence d'une séquence de référence du génome de blé, nous avons utilisé les 6000 EST de blé qui ont été cartographiées sur les bins de délétion (Qi et al. 2004) afin d'identifier les régions synténiques avec le génome du riz, utilisé ici comme modèle. Une EST correspond à une séquence de 300 à 500 nucléotides issue du séquençage partiel d'un ADNc cloné, donc caractéristique d'un gène exprimé. Une recherche similarité a été réalisée par BLASTX entre les 6000 EST de blé dont la localisation chromosomique est connue et les produits de gènes du riz. Ainsi, il a été possible d'identifier des blocs de synténie c'est-à-dire des régions contiguës de colinéarité entre deux génomes (Figure 9). Ainsi, les gènes portés par le chromosome 3B du blé sont majoritairement portés par le chromosome 1 du riz (et le chromosome 2 de *Brachypodium*), révélant leur origine commune. Toutefois, cette étude a permis d'évaluer que 46% des gènes cartographiés chez le blé présentent une plus forte similarité (meilleur "hit" de BLAST) avec un gène porté par une région non synténique du génome de riz. Ceci suggère que, bien que la macro-colinéarité entre les génomes du blé et des autres *Poaceae* soit fortement conservée, la conservation est moindre à l'échelle de la micro-colinéarité.

2. Sélection de deux sets de gènes de blé synténiques et non synténiques avec les génomes des espèces modèles

Au cours de ce stage, nous avons focalisé notre étude sur les séquences complètes de 13 grands contigs de BAC (de 500 kb à 3,2 Mb) représentant 18 Mb répartis le long du chromosome 3B du blé (Choulet et al. 2010) (Figure 10). Leur annotation avait révélé la présence de 199 régions géniques non dérivées d'éléments transposables. L'annotation a été réalisée en combinant les résultats de recherche de similarités avec des ADNc, des EST, des unigènes (clusters d'EST) ainsi qu'avec les séquences protéiques des gènes annotés dans les génomes apparentés et ceux présents dans les banques de séquences plus généralistes (SwissProt, NR). Parmi les 199 régions présentant une similarité avec des gènes codant des protéines, 51 ont été classées comme pseudogènes et fragments de gènes, car ne présentant pas une structure fonctionnelle : codons stop dans l'ORF, décalage du cadre de lecture, troncature importante. Ils n'ont pas été pris en compte dans les analyses réalisées au cours du stage. En effet, l'un des objectifs du stage était d'étudier l'aspect fonctionnel des gènes non synténiques ainsi que leur expression. Afin de réduire le set de gènes à analyser, il est apparu approprié de filtrer ceux apparaissant comme inactifs sur la base de leur séquence uniquement.

A

Gènes non synténiques	Plus proche homologue chez le riz	% d'identité
ctg0011b.00010	Os08g20200	73
ctg0011b.00040	Os08g20200	69
ctg0011b.00060	Os12g13030	71
ctg0011b.00070	Os12g13030	72
ctg0011b.00100	Os07g42354	71
ctg0011b.00110	Os12g13030	72
ctg0011b.00120	Os07g42354	72
ctg0011b.00140	Os10g32300	86
ctg0011b.00180	Os02g26290	61
ctg0011b.00190	Os03g01270	78
ctg0011b.00200	Os03g01270	80
ctg0011b.00210	Os07g02140	61
ctg0011b.00230	Os11g38010	64
ctg0011b.00250	Os11g38650	81
ctg0011b.00340	Os08g06070	85
ctg0079b.00010	Os04g16680	94
ctg0091b.00060	Os12g26470	44
ctg0091b.00080	Os03g30950	81
ctg0091b.00150	Os08g09390	56
ctg0091b.00160	Os08g09390	61
ctg0382b.00140	Os09g06620	63
ctg0464b.00020	Os11g05730	99
ctg0464b.00070	Os12g42140	35
ctg0464b.00080	Os10g04700	41
ctg0464b.00160	Os08g24370	39
ctg0464b.00180	Os03g25330	78
ctg0464b.00210	Os03g63850	61
ctg0616b.00200	Os08g21840	83
ctg0661b.00020	Os03g04650	68
ctg0661b.00030	Os09g16380	74
ctg0661b.00040	Os03g01270	80
ctg0661b.00050	Os03g04650	74
ctg0954b.00010	Os10g32080	63
ctg0954b.00040	Os02g04260	58
ctg0954b.00050	Os10g38600	70
ctg0954b.00070	Os02g09540	48
ctg0954b.00090	Os10g17960	59
ctg0954b.00130	Os11g28104	73
ctg0954b.00170	Os08g09570	52
ctg0954b.00180	Os02g29380	55
ctg0954b.00200	Os08g36320	80
ctg0954b.00215	Os11g40590	69
ctg0954b.00220	Os03g13220	92
ctg0954b.00260	Os12g44360	82
ctg0954b.00270	Os05g50360	89
ctg0954b.00280	Os02g04250	44
ctg0954b.00300	Os05g01200	54
ctg0954b.00340	Os04g02150	78
ctg0954b.00360	Os05g46520	62
ctg0954b.00380	Os03g24690	51
ctg0954b.00460	Os05g12040	69
ctg0954b.00470	Os12g03750	42
ctg0954b.00480	Os12g03740	46

B

Gènes synténiques	Plus proche homologue chez le riz	% d'identité
ctg0005b.00010	Os01g54670	91
ctg0005b.00020	Os01g54630	83
ctg0005b.00030	Os01g54620	91
ctg0005b.00040	Os01g54600	64
ctg0005b.00050	Os01g54590	91
ctg0005b.00060	Os01g54580	81
ctg0005b.00070	Os01g54570	56
ctg0005b.00090	Os01g54560	88
ctg0005b.00100	Os01g54550	73
ctg0011b.00080	Os01g02300	63
ctg0011b.00090	Os01g02290	65
ctg0011b.00320	Os01g02200	90
ctg0079b.00020	Os01g63290	85
ctg0079b.00030	Os01g63280	74
ctg0079b.00040	Os01g63270	93
ctg0079b.00050	Os01g63270	93
ctg0079b.00060	Os01g63260	73
ctg0079b.00070	Os01g63250	77
ctg0079b.00080	Os01g63230	84
ctg0079b.00090	Os01g63220	79
ctg0079b.00100	Os01g63210	80
ctg0079b.00110	Os01g63200	75
ctg0079b.00120	Os01g63190	80
ctg0079b.00130	Os01g63160	62
ctg0091b.00010	Os01g68510	76
ctg0091b.00020	Os01g68500	67
ctg0091b.00040	Os01g68480	76
ctg0091b.00050	Os01g68450	78
ctg0091b.00090	Os01g68440	65
ctg0091b.00100	Os01g68390	82
ctg0091b.00110	Os01g68380	71
ctg0091b.00170	Os01g68370	60
ctg0091b.00190	Os01g68330	91
ctg0091b.00200	Os01g68324	86
ctg0382b.00050	Os01g48950	36
ctg0382b.00080	Os01g48874	58
ctg0382b.00090	Os01g48874	71
ctg0382b.00100	Os01g48850	84
ctg0382b.00130	Os01g48830	68
ctg0382b.00160	Os01g48800	84
ctg0464b.00030	Os01g64650	86
ctg0464b.00040	Os01g64660	89

Gènes synténiques	Plus proche homologue chez le riz	% d'identité
ctg0464b.00090	Os01g64680	83
ctg0464b.00100	Os01g64690	98
ctg0464b.00110	Os01g64700	84
ctg0464b.00130	Os01g64720	85
ctg0464b.00150	Os01g64730	61
ctg0464b.00170	Os01g64750	84
ctg0464b.00200	Os01g64760	91
ctg0464b.00230	Os01g64770	83
ctg0464b.00250	Os01g64780	94
ctg0464b.00260	Os01g64790	59
ctg0464b.00310	Os01g64820	90
ctg0528b.00010	Os01g63510	84
ctg0528b.00020	Os01g63540	92
ctg0528b.00030	Os01g63580	91
ctg0528b.00040	Os01g63620	81
ctg0616b.00190	Os01g48710	64
ctg0616b.00210	Os01g48720	77
ctg0616b.00230	Os01g48760	90
ctg0661b.00070	Os01g71380	62
ctg0954b.00020	Os01g02880	90
ctg0954b.00060	Os01g02884	91
ctg0954b.00080	Os01g02890	87
ctg0954b.00140	Os01g02900	60
ctg0954b.00160	Os01g02920	71
ctg0954b.00230	Os01g02930	57
ctg0954b.00240	Os01g02930	61
ctg0954b.00250	Os01g02940	65
ctg0954b.00290	Os01g03020	87
ctg0954b.00310	Os01g03030	76
ctg0954b.00320	Os01g03040	71
ctg0954b.00330	Os01g03040	75
ctg0954b.00370	Os01g03050	82
ctg0954b.00390	Os01g03060	67
ctg0954b.00420	Os01g03070	80
ctg0954b.00430	Os01g03080	60
ctg0954b.00440	Os01g03090	79
ctg0954b.00450	Os01g03100	84
ctg0954b.00490	Os01g03110	77
ctg1035b.00020	Os01g34330	76

Table2: Homologie entre les gènes du blé et du riz.
Liste des 53 gènes de blé non synténiques avec le génome du riz (gauche) et des 81 synténiques (droite). Pour chaque gène étudié, l'homologue de plus forte similarité chez le riz à été identifié par un BLASTP. Le pourcentage d'identité indiqué correspond au niveau de conservation des séquences protéiques.

Une recherche de similarité par BLASTP a été réalisée pour les 148 produits de gènes potentiellement fonctionnels avec l'ensemble des protéines annotées chez le riz afin d'identifier les couples d'orthologues (similarité de séquences et localisation en position de synténie) : ces gènes ont été appelés "gènes synténiques" dans ce rapport. De la même manière, les gènes de blé présentant une plus forte similarité avec des gènes de riz (Table 2) et de *Brachypodium* portés par des régions non synténiques ont été sélectionnés et appelés "gènes non synténiques" dans ce rapport. Ainsi, respectivement 81 et 61 gènes synténiques et non synténiques ont été identifiés. Parmi les 61 gènes non colinéaires, 8 présentent la plus forte similarité avec des gènes localisés sur le chromosome 1 du riz (l'orthologue du 3B) et seraient donc issus de duplications/translocations intra-chromosomiques. Ils ont été éliminés de l'analyse pour se focaliser sur les duplications inter-chromosomiques. Ainsi, cette étude a porté sur l'analyse comparative de deux échantillons composés, d'une part, de 53 gènes non synténiques et, d'autre part, de 81 gènes synténiques.

3. Localisation du locus ancestral ayant été dupliqué

Partant des résultats obtenus concernant les relations d'orthologie entre les chromosomes du blé, du riz et de *Brachypodium*, l'identification *in silico* du chromosome de blé portant la copie ancestrale des gène non synténiques a été possible. Chaque gène non synténique n'est pas spécifique du génome de blé. En effet, une recherche de similarité avec les génomes de riz de *Brachypodium* a montré une conservation avec des loci non synténiques du chromosome 3B. Cette similarité entre gènes non colinéaires suggère fortement des relations de paralogie, c'est-à-dire indiquant que l'homologie de séquences est issue de la duplication d'un gène qui s'est produite spécifiquement dans la lignée des Triticeae après sa divergence avec les espèces voisines. La localisation de se gène ayant été dupliqué peut ainsi être prédite en combinant les relations de synténie établie à l'échelle des génomes complets (cf paragraphe précédent) avec la plus forte similarité retrouvée pour chaque gène. Par exemple, le gène ctg09540.00030.1 (porté par le chromosome 3B) présente une plus forte similarité avec Os05g01200.1 et Bradi2g40020.1. Ces derniers sont colinéaires entre riz et *Brachypodium* et sont synténiques avec les chromosomes 1A/1B/1C du génome de blé. Nous suggérons donc que le locus ancestral porté par ce chromosome a été dupliqué sur le 3B (Annexe 1).

La répartition des copies ancestrales des gènes non synténiques étudiés est homogène sur l'ensemble du génome du blé. Ceci révèle que les cibles des événements de duplication de gènes sont aléatoirement répartis dans le génome et qu'aucun chromosome ou aucune région particulière n'aurait été la cible privilégiée de duplications

A Gènes non synténiques

Famille 1	ctg0011b.00060	Ctg0011b.00070	Ctg0011b.00110
Famille 2	Ctg0011b.00100	Ctg0011b.00120	
Famille 3	Ctg0011b.00190	Ctg0011b.00200	
Famille 4	Ctg0091b.00150	Ctg0091b.00160	
Famille 5	Ctg0661b.00020	Ctg0661b.00050	

B Gènes synténiques

Famille 1	ctg0005b.00040	ctg0079b.00080	
Famille 2	ctg0011b.00320	ctg1035b.00020	
Famille 3	ctg0079b.00040	ctg0079b.00050	
Famille 4	ctg0079b.00100	ctg0528b.00020	
Famille 5	ctg0079b.00110	ctg0079b.00120	
Famille 6	ctg0091b.00020	ctg0954b.00440	
Famille 7	ctg0382b.00080	ctg0382b.00090	
Famille 8	ctg0382b.00130	ctg0954b.00290	
Famille 9	ctg0954b.00140	ctg0954b.00160	ctg0954b.00230 ctg0954b.00240 ctg0954b.00250
Famille 10	ctg0954b.00320	ctg0954b.00330	

Table 3 : Liste des gènes dupliqués au sein de l'échantillon étudié.
Les séquences codantes des 53 et 81 gènes synténiques et non synténiques ont été comparées entre elles par BLASTN. Le programme blastclust.pl a permis de regrouper les gènes en familles dupliquées sur la base des critères suivants : plus de 90% d'identité sur au moins 40% de leur longueur.

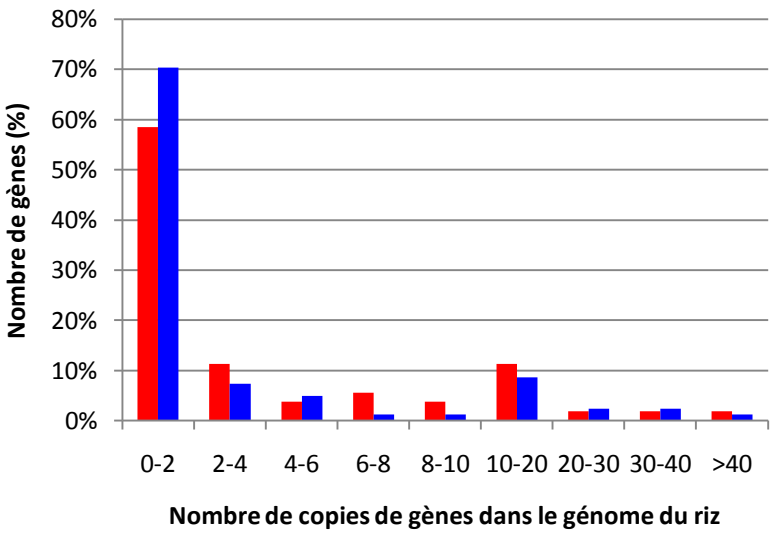


Figure 11 : Distribution des 134 meilleurs homologues chez le riz des gènes de blé synténiques (bleu) et non synténiques (rouge) en fonction du nombre de copies dupliquées au sein du génome de riz.
Pour chaque gène étudié, l'homologue de plus forte similarité chez le riz à été aligné avec l'ensemble des protéines du génome du riz par BLASTP. Blastclust.pl a permis de classer les protéines présentant plus de 40% sur au moins 70% de leur longueur au sein d'une même famille dupliquée et de compter le nombre de membres de chaque famille.

4. Caractère intrinsèquement répété des gènes étudiés : dupliqués en tandem ou appartenant à de grandes familles multigéniques

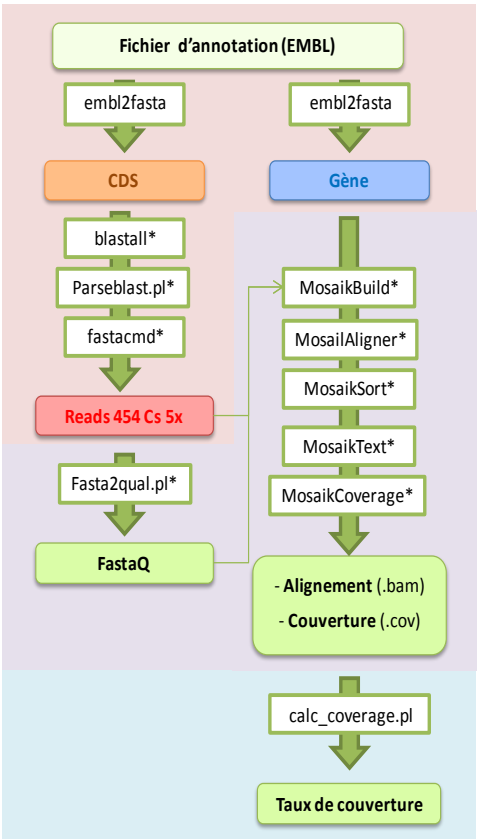
L'un des objectifs de cette étude était d'identifier le nombre et la localisation des copies récemment dupliquées de chaque gène candidat. Toutefois, il est déjà connu qu'au sein des génomes certains gènes sont redondants : les gènes dupliqués en tandem, et les gènes appartenant à des familles géniques largement représentées comme, par exemple, les gènes de protéines ribosomiques, les histones etc. Ainsi, préalablement à cette étude, nous avons entrepris de caractériser de type de redondance génique pour chaque des candidats.

Pour identifier les gènes ayant été dupliqués localement (en tandem), chacune des CDS des 53 et 81 gènes synténiques et non synténiques a été alignée par BLASTN sur l'ensemble des CDS du set étudié. Puis les CDS partageant une identité supérieure à 90% sur une longueur d'au moins 40% de leur taille ont été regroupés en familles par le programme blastclust.pl.

Ainsi, il est apparu que 20% des CDS non synténiques et 28% des CDS synténiques partage avec au moins une autre CDS 90% ou plus d'identité. De plus, il apparaît que l'ensemble des gènes identifiés comme similaires sont localisés au sein du même contig ou au sein de contigs physiquement très proche, confirmant ainsi qu'il ont probablement subit une duplication en tandem (Table 3).

Les génomes de Poaceae contiennent une proportion significative de gènes appartenant à de grandes familles fonctionnelles et sont ainsi présents à un grand nombre de copies dans la plupart des génomes. Ces gènes peuvent biaiser les résultats de cette étude et devaient donc être identifiés. Ne disposant pas de la composition du génome du blé, nous avons basé cette étude sur le génome d'*Oryza sativa*. En utilisant BLASTP, les meilleurs homologues chez le riz des 134 gènes étudiés ont été comparés à l'ensemble des gènes annotés dans ce même génome afin de dénombrer le nombre de membres de la même famille. Grace au programme *ParseBlast.pl*, seuls les protéines s'alignant avec plus de 40% d'identité sur au moins 70% de leur longueur ont été considérées comme homologues et, donc, appartenant à la même famille. Les résultats indiquent que seulement 47% des gènes ainsi étudiés sont en copie unique au sein du génome du riz. En revanche, 34% des gènes étudiés font partie de familles multigéniques (considérant un seuil d'au moins 3 membres). On retrouve dans cet échantillon des gènes codant une histone H3, une F-box, ou encore une peroxidase. De façon intéressante, la distribution du nombre de copies identifiées chez le riz (Figure 11) est différente pour les gènes synténiques et non synténiques : respectivement 41% et 29% des gènes non synténiques et synténiques appartiennent à des

A



B

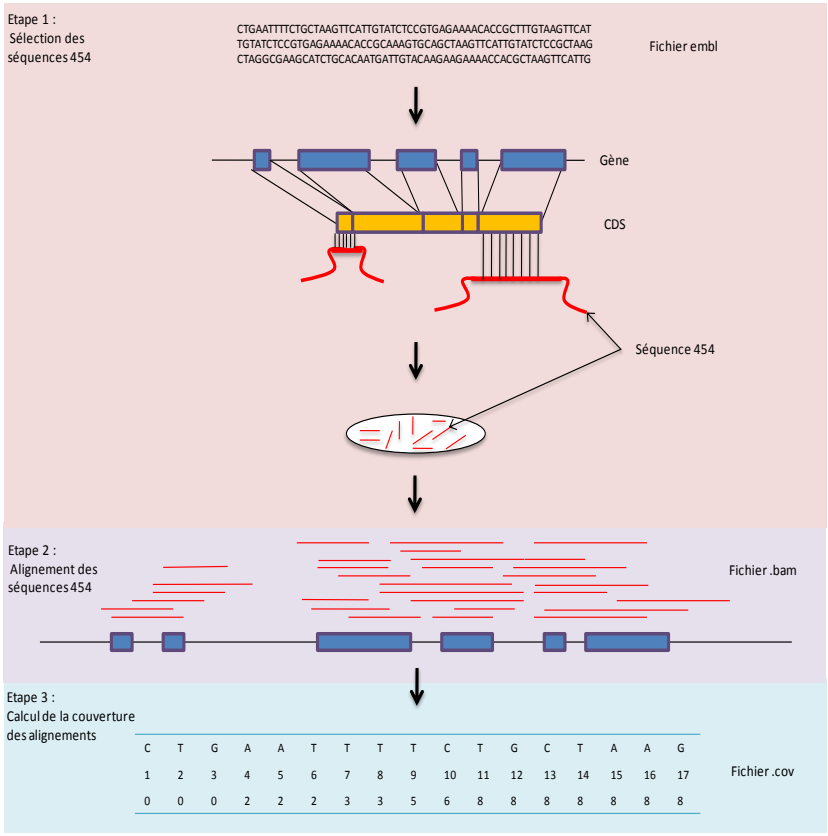


Figure 12 : Modèle conceptuel de l'approche bioinformatique déterminant le nombre de copies des gènes étudiés au sein du génome sans étape de clustering".

(A) Suivi du flux des données.
*Programmes ayant été intégrés dans le pipeline BlastMap.

(B) Représentation schématique du processus d'analyse. A partir d'un fichier EMBL, la séquence du gène et la séquence codante (CDS) sont extraites. Puis, un BLAST est réalisé entre la CDS et les séquences 454 du set CS5x. Les séquences s'alignant sur la CDS selon les critères choisis sont sélectionnées (étape 1), puis, à l'aide de la suite Mosaik, les séquences 454 sont alignées sur la séquence du gène (étape 2). Un fichier détaillant le taux de couverture pour chaque position nucléotidique est ainsi généré (étape 3). Le programme *calc_coverage.pl* a été développé pour calculer automatiquement le taux de couverture moyen par exon à partir de ce fichier.

familles de gènes répétées chez le riz. Cette différence suggère que les gènes appartenant à des familles à grand nombre de copies ont une tendance accrue à être dupliqué.

5. Etablissement de méthodes de détermination du nombre de copies des gènes candidats au sein du génome hexaploïde du blé

Bien qu'aucune séquence de référence du génome de blé ne soit disponible, un effort de séquençage "génome entier" (Whole Genome Shotgun, WGS) a été produit en 2010 dans le cadre d'un projet britannique (http://www.cerealsdb.uk.net/). Plus de 76 Gb, représentant près de 200 millions de lectures Roche-454 de 380 pb en moyenne, ont ainsi été générées à partir du génome hexaploïde de la variété Chinese Spring (variété de référence internationale). Cet échantillon représente donc une couverture de 5x du génome (70 Gb/17 Gb). Par simplicité, il sera nommé "CS5x" dans ce rapport. Si cette couverture est très insuffisante pour assembler une séquence de référence de qualité, elle est suffisante pour identifier des séquences partielles de gènes d'intérêt. Sur la base de ces séquences, deux stratégies différentes ont été mises en place pour estimer le nombre de copies des gènes 134 étudiés ici. La première approche (Figure 12) se base sur le taux de couverture moyen, de chaque gène candidat, obtenu en alignant toutes les lectures CS5x qui présentent une similarité significative avec le gène étudié. La seconde approche (Figure 13) est basée sur le "clustering" des séquences ainsi alignées afin d'assembler les différentes copies du même gène en contigs.

a) Détermination du nombre de copies de gènes par l'évaluation du taux de couverture dans les séquences CS5x

1) Description de l'approche

<u>Etape 1 : Sélection des séquences 454 à aligner :</u>

À partir des fichiers contenant l'annotation au format EMBL des 13 contigs de BAC séquencés précédemment, la séquence nucléotidique de la CDS (séquence codante ; exons seulement) de l'ensemble des 134 gènes étudiés a été extraite à l'aide du programme *embl2fna.pl*. Une recherche de similarité a ensuite été réalisée par BLASTN sur l'ensemble séquences 454 du set CS5x. A cette étape, les parties introniques ont volontairement été exclues de la recherche de similarité. En effet, les introns contiennent fréquemment des éléments transposables (notamment de type MITE) qui, au vu de la forte proportion en éléments répétés dans le génome du blé, auraient été alignés de façon aspécifique avec d'autres séquences de la même famille d'éléments.

Le résultat de BLASTN a ensuite été filtré avec *ParseBlast.pl* en testant plusieurs valeurs de seuil pour l'identité et la longueur de l'alignement afin d'optimiser l'approche pour ne conserver que les séquences 454 alignées de façon significative : identité minimale égale à 80% ou 90% ; longueur minimale de l'alignement : 50 nucléotides ou 40% de la taille de la séquence 454. Enfin, le programme *fastacmd* (suite ncbi-blast) a été utilisé pour extraire les séquences 454 CS5x ayant été ainsi filtrées à partir de l'alignement et pour ainsi créer un fichier multiFASTA.

Étape 2 : Alignement des séquences génomiques CS5x sur leur séquence génique de référence

Bien que la CDS ait été utilisée pour la recherche de similarité, la séquence de référence à utiliser lors de cette seconde étape doit être une séquence génomique (exon+intron+UTR). Nous avons donc extrait des séquences des 13 contigs de BAC, les séquences nucléotidiques des gènes (exon+intron) ainsi que les 200 pb flanquants en 5' et 3' à l'aide du programme subseq.pl. L'étape qui a consisté à aligner les séquences 454 identifiées par BLAST sur ces séquences de référence (anglais : "read mapping") a été réalisée à l'aide du programme *Mosaik* (Barnett et al. 2011).

MosaikBuild a été le premier sous programme à être utilisé. Son rôle a été de convertir les séquences de référence et les séquences à aligner au format FASTA en format de fichier binaire spécifique à *Mosaik* (création d'un index à partir des séquences). Cette étape nécessite l'attribution d'une valeur de qualité affectée à chaque nucléotide des séquences. Ne disposant pas de cette information, nous avons attribué arbitrairement une valeur de 40 (maximum) à chaque nucléotide et avons ainsi créé un ficher au format FastaQ à l'aide au programme fasta2qual.pl.

Puis chaque séquence 454 a été ensuite alignée, indépendamment les unes des autres, sur la séquence de référence par le programme *MosaikAligner*. Seuls les alignements entre la séquence de référence et la lecture 454 présentant une homologie d'au moins 90% ont été retenu.

Puis le programme *MosaikSort* a permis de regrouper les alignements individuels de chaque séquence 454 sur la séquence de référence en alignement commun. Enfin par l'intermédiaire de deux programmes, *MosaikText* et *Samtools* (Li et al. 2009), il a été possible de convertir le fichier binaire d'alignement de *Mosaik* en un format BAM permettant la visualisation des alignements avec *ARTEMIS* par exemple (Carver et al. 2008).

Étape 3 : Calcul du taux de couverture de chaque gène étudié dans le set CS5x :

L'alignement des séquences 454 sur une séquence de référence permet d'évaluer pour chaque nucléotide de la séquence de référence, le nombre de lectures 454 qui le couvre. Cette opération a été réalisée par l'intermédiaire du programme *MosaikCoverage*. Ce programme assigne à chaque nucléotide de la séquence de référence une valeur représentant le nombre de fois que celle-ci est couverte par une séquence 454 alignée.

Toutefois, seule la couverture des exons est significative puisque, bien que les alignements des séquences 454 couvre une partie des introns, le taux de couverture des introns est nettement sous estimée puisque seule une similarité avec la séquence codante a été recherchée initialement. Ainsi, le programme *calc_coverage.pl* a été développé dans le but de calculer précisément le taux de couverture en tenant compte des positions des exons et des introns.

Ce programme s'appuie sur les positions identifiées sur les 13 contigs de BAC annotés. Il a donc fallu recalculer ces positions exon/introns sur chaque séquence de référence. Le programme *exon_recalc.pl* a donc été développé afin d'automatiser ce calcul pour les 134 gènes étudiés.

Ainsi grâce à ce processus, la couverture moyenne de chaque exon et de chaque gène ont été calculée. D'après l'estimation à 5x du taux de couverture du génome (CS5x), une simple division par 5 permet ainsi d'estimer le nombre de copies de chaque gène au sein du génome hexaploïde.

2) Les limites de l'approche : Les hétérogénéités locales du taux de couverture de CS5x

Les 76 Gb séquencés couvrant le génome hexaploïde représentent une couverture de 5x en moyenne. Mais des variations locales du taux de couverture étaient attendues, et leur ampleur peut grandement influencer les résultats. En effet, certaines copies de gènes peuvent ne pas avoir été séquencées ou, au contraire, être couvertes plus de 10x.

Dans le but de contrôler l'intensité de ces variations le long du génome, nous avons utilisés les séquences des 53 gènes non synténiques candidats (exons+introns+UTR) comme "séquences contrôles". Une unique séquence de référence chimérique a donc été créée en concaténant ces 53 séquences génomiques du chromosome 3B, séparées par des trous matérialisés par 100 bases "N". Les éléments transposables insérés dans les introns et les régions UTR ont ensuite été masqués à l'aide du programme *RepeatMasker* (Jiang et al. 2008). Cette opération consiste a transformer les nucléotides des TE en minuscule. Ils seront ainsi ignorés par BLAST. Une recherche de similarité dans la banque de séquences CS5x a ensuite été réalisée par BLASTN.

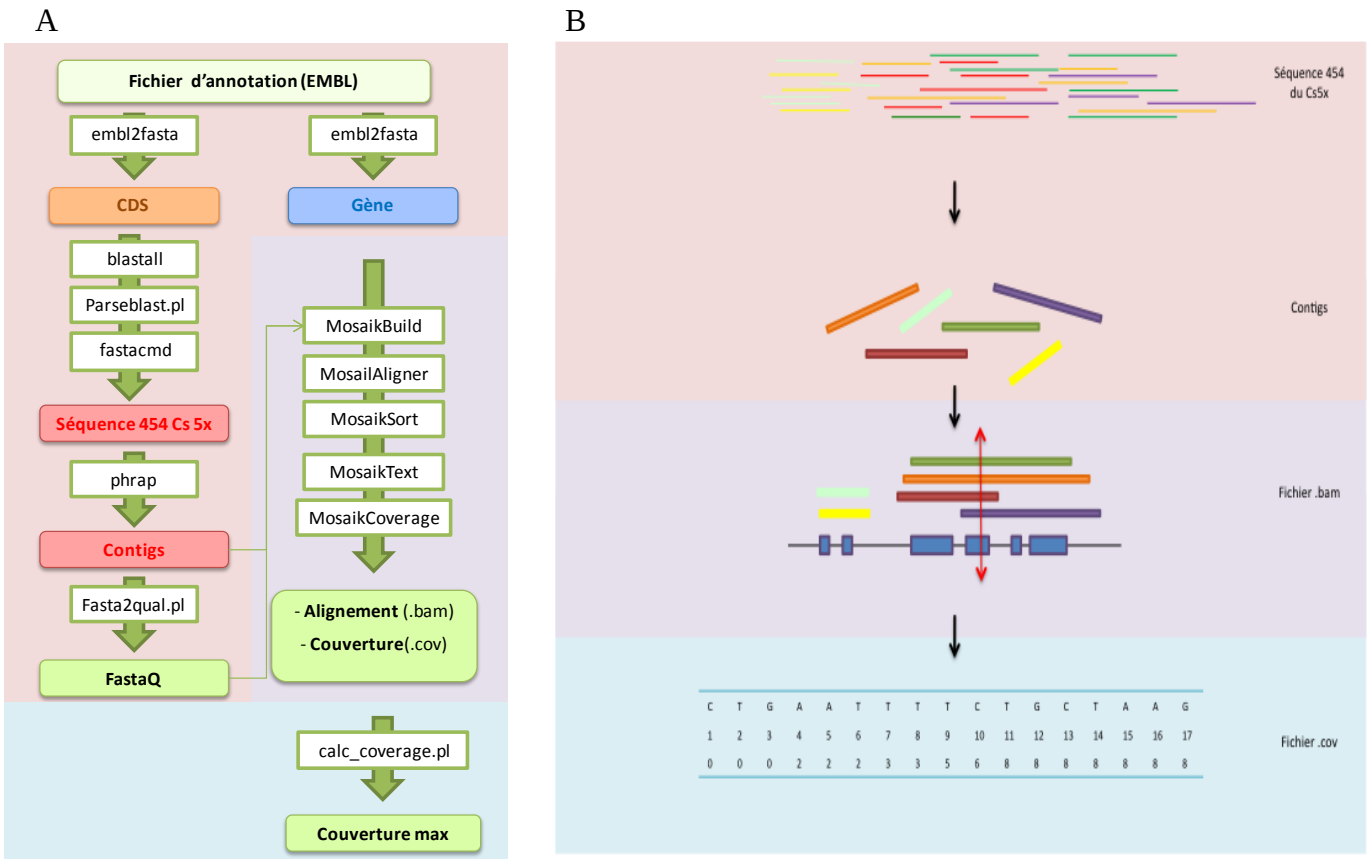


Figure 13 : Modèle conceptuel de l'approche bioinformatique déterminant le nombre de copies des gènes étudiés au sein du génome avec une étape de "clustering".

(A) Suivi du flux des données. (cf figure précédente)
L'étape de "clustering" est réalisée par l'assembleur Phrap.

(B) Représentation schématique du processus d'analyse. Les différentes couleurs des séquences représentent les séquences provenant des différentes copies divergentes d'un même gène. Le programme Phrap assemble les séquence 454 identiques en contigs. A l'aide du programme Mosaik les contigs sont alignées sur la séquence du gène de référence. Un fichier détaillant le taux de couverture pour chaque position nucléotidique est ainsi généré et analysé avec *calc_coverage.pl* qui permet de déterminer la couverture maximale (flèche rouge) correspondant au nombre maximum de copies divergentes du même gène identifiées.

L'identité de séquence minimale considérée pour le filtrage des alignements a été de 98%, c'est-à-dire une identité quasi-parfaite, aux erreurs de séquençage près (pouvant atteindre 2%), afin de n'identifier que les séquences provenant du gènes 3B et non des autres copies légèrement divergentes. Après ce filtrage, le programme Mosaik a été utilisé pour aligner les séquences préalablement identifiées sur la molécule chimérique de référence.

Enfin, à partir du taux de couverture décrit pour chaque nucléotide (par *MosaikCoverage*), le script *cov.pl* a été développé pour calculer automatiquement un taux de couverture moyen et un écart type, permettant d'estimer l'hétérogénéité du séquençage "shotgun". Ce script permet d'exclure toutes les valeurs de couverture affectées aux nucléotides masqués (TE) et aux nucléotides N (ayant une couverture de 0) afin d'éviter tout biais possible.

Ainsi sur l'ensemble de la séquence de référence la couverture moyenne obtenue a été de 6.1x+/-4 avec des variations allant de 0x à 26x pour certaines régions. Ce résultat confirme qu'une hétérogénéité importante du taux de couverture existe localement. L'extrapolation du taux de couverture moyen par gène peut donc biaiser l'estimation du nombre de copies identifiées par cette approche.

b) Détermination du nombre de copies de gènes par "clustering" des lectures 454

1) Description de la méthode

À la suite de l'évaluation du WGS en 5x montrant des grandes variations de l'homogénéité du séquençage, une approche tentant de corriger ce biais a été entreprise. Le principe de cette stratégie est de regrouper les séquences 454 alignées sur les séquences de référence sur la base de leur identité (="clustering") (Figure 13). Ce clustering a permis de générer des contigs de séquences (consensus) qui ont ensuite été alignés sur leur gène de référence pour estimer un nombre de copies légèrement divergentes de chaque gène. Les critères d'alignement entre séquences 454 et gène de référence ont été les mêmes que précédament. Puis l'assembleur *Phrap* (Gordon et al. 1998) a été utilisé pour créer des contigs à partir des séquences 454 qui présentent une zone de chevauchement parfaitement identique. Le pipeline *BlastMap* et *calc_coverage.pl* ont été ensuite utilisés pour aligner les contigs ainsi obtenus et calculer un taux de couverture.

2) Mode de calcul du nombre de copies

Les contigs ont été obtenus à partir des séquences 454 ayant des similitudes avec la CDS. Par conséquent, les grands introns ne sont pas couverts par des séquences 454 et les contigs issus de l'assemblage peuvent ne couvrir que partiellement un gène. Ceci signifie que, pour une même

copie de gène, plusieurs contigs recouvrant partiellement le même gène peuvent être assemblés (Figure 13b). Le nombre de contigs totaux ne correspond donc pas au nombre de copies de gènes. Par ailleurs, la visualisation des alignements avec *ARTEMIS* a montré que chaque exon d'un gène n'est pas toujours couvert à un niveau identique. Ainsi, l'estimation la plus judicieuse du nombre de copies de gènes présents dans le génome du blé, équivaut au nombre maximum de contigs alignés sur le même exon.

c) Les limites de l'approche par clustering

Il existe plusieurs raisons expliquant la variation du nombre de contigs mappés par exon. Premièrement la variation du WGS peut expliquer que, même avec une couverture moyenne de 5x, certaines portions de gènes ne soient pas séquencées. Ensuite il est possible que plusieurs copies du même gène soient très peu divergentes, cas des gènes ayant été dupliqués très récemment. Dans ce cas précis, les séquences 454 issues de différentes copies du même gène ont été assemblées en un seul contig. Enfin certaines copies de gènes peuvent avoir perdu des exons.

6. Développement d'un pipeline bioinformatique : *BlastMap*

La stratégie de bioinformatique consistant à aligner les séquences 454 sur une séquence de référence a largement été utilisée au cours de cette étude. Composé de 11 étapes successives, le cheminement de cette analyse s'avère long et répétitif (9 des principaux programmes du pipeline sont présentés dans la figure 12a). Ainsi, par soucis d'optimiser cette stratégie et en prévision des utilisations futures (le chromosome 3B complet est en cours de séquençage), le pipeline BlastMap a été développé au cours du stage. Ce pipeline permet, via une ligne de commande unique, de lancer toutes les différentes étapes de l'analyse successivement en gérant automatiquement les fichiers d'entrée et de sortie des différents programmes.

Le langage informatique utilisé pour l'écriture du script a été le langage Perl avec l'utilisation des modules BioPerl (Stajich 2007). Un fichier de configuration au format xml a été créé parallèlement. Il détaille l'ensemble des options paramétrables pour les différents programmes exécutés au cours de l'analyse et permet donc de faire varier ces paramètres sans modifier le code du programme (Annexe 5).

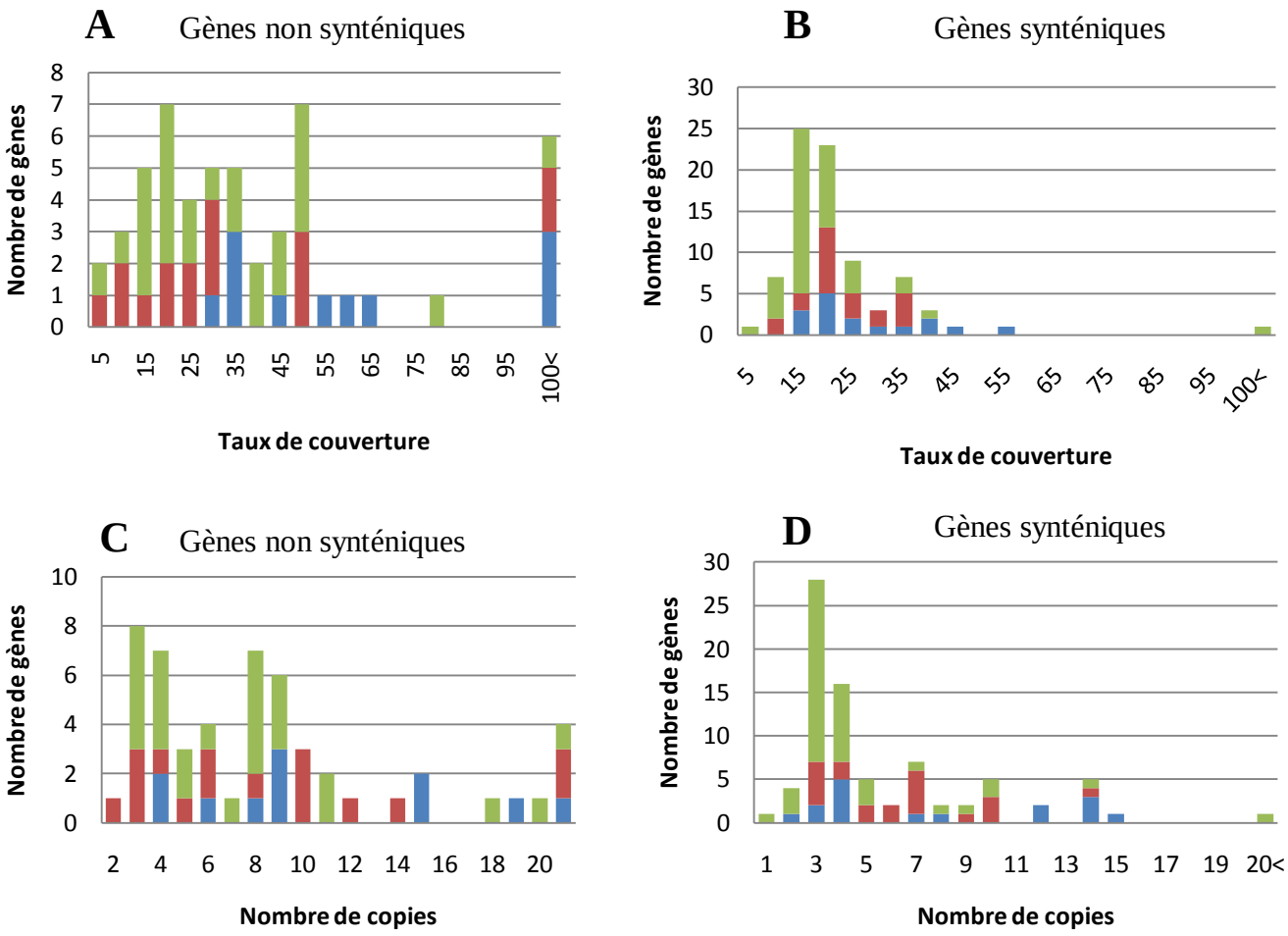


Figure 14 : Distribution du taux de couverture et du nombre de copies des gènes non synténiques (A-C) et synténiques (B-D)

A et B. Distribution du nombre de copies de gènes estimé sur la base du taux de couverture moyens obtenus pour les gènes synténiques et non synténiques.
C et D. Distribution du nombre de copies de gènes estimés après "clustering" des séquences 454 identiques.

Pour chaque histogramme, les gènes identifiés comme dupliqués en tandem sont représentés en bleu, ceux appartenant à des familles de gènes dupliqués (>2 copies dupliquées chez le riz) sont représentés en rouge. Les autres gènes sont représentés en vert.

7. Estimation du nombre de copies des gènes étudiés

a) Distribution du nombre de copies des gènes synténiques et non synténiques

Les distributions du nombre de copies des 134 gènes candidats, estimées par les 2 approches décrites précédemment, sont présentées à la figure 14 pour les gènes non synténiques (A-C) et les gènes synténiques (B-D).

Pour les deux catégories de gènes, les deux types d'approches donnent des résultats très similaires (Annexe 2). Il y a seulement 10 gènes pour lesquelles une grosse différence existe (considérant un seuil d'au moins 5 copies de différence).

Par ailleurs, les distributions du nombre de copies observées apparaissent très différentes pour les gènes synténiques par rapport aux non synténiques. En effet, pour des gènes synténiques, la distribution présente un pic autour de 3 à 4 copies, suggérant la présence attendue de 3 copies homéologues A-B-D. Ce pic représente 54% des gènes synténiques. En revanche, pour les gènes non synténiques, la distribution montre en général que le nombre de copies de gènes est plus élevé, suggérant que ces gènes ont effectivement subit des duplications récentes. Trois pics sont visibles sur cette distribution : le premier autour de 3-4 copies, le second à 8-9 copies, et un dernier supérieur à 18 copies. Au total, 83% des gènes non synténiques présentent un nombre de copies supérieur à trois. D'autre part, les gènes préalablement caractérisés comme dupliqués en tandem ou appartenant à de grandes familles multi-géniques, sont majoritairement représentés dans les classes à nombre de copies élevé. Ainsi, une partie des gènes ayant subit des duplications interchromosomiques sont des gènes qui était déjà largement dupliqués au sein du génome.

b) Les fonctions de gènes ayant tendance à être plus fréquemment dupliquées

Pour déterminer la fonction des gènes étudiés, une recherche de domaines protéiques *Pfam* a été réalisée à l'aide de *HMMER* (Finn et al. 2011). Ainsi, 342 domaines ont été identifiés pour les 134 séquences protéiques. Pour réaliser une classification des fonctions en catégories, nous nous sommes appuyés sur la "Gene Onthology" (www.geneontology.org). Pour cela, des termes "GO" ont été attribués à chacun des gènes en fonction des domaines protéiques identifiés grâce à *Pfam2Go* (Bateman et al. 1999). Enfin, le logiciel *GOBO* a permis d'attribuer une catégorie fonctionnelle adéquate en fonction des termes GO affectés précédemment. Parmi les "fonctions moléculaires" identifiées (Figure 15), les gènes synténiques et non synténiques sont tous les deux majoritairement représentés par les fonctions protéiques de liaison à un substrat et d'activité catalytique. Cette étude ne révèle aucun biais fonctionnels entre les gènes synténiques

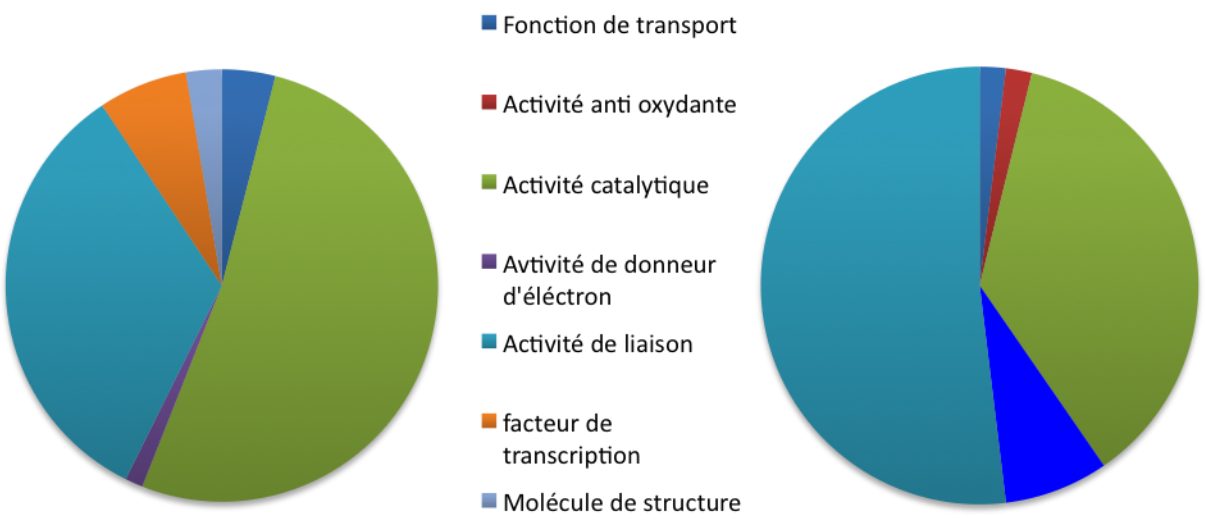


Figure 15 : Fonctions moléculaires attribuées aux gènes synténiques (A) et non synténiques (B).

La classification par catégorie fonctionnelle a été réalisée par recherche de domaines Pfam dans les séquences des produits des 134 gènes étudiés. Puis, des termes de la "Gene Ontology" ont été attribués sur la base des domaines Pfam identifiés.

et non synténiques. Les duplications inter-chromosomiques ne semblent donc pas affecter majoritairement certaines catégories de fonctions. Toutefois, l'échantillon étudié est peut-être trop petit pour être affirmatif à ce sujet.

1. Amorce de la détermination du nombre de copies de gènes par une approche de biologie moléculaire

Afin de localiser les 53 gènes non synténiques identifiés grâce à l'approche bioinformatique, une étude par amplification PCR a été entreprise. Celle-ci consiste en l'identification par présence/absence de chaque gène sur les bras de chromosomes triés du génome de blé.

a) Dessin des amorces

Au total 53 couples d'amorces ont été définis pour l'ensemble des gènes non synténiques. Les critères du design d'amorce utilisés sont : une taille d'amorce d'environ 20 pb, une taille d'amplicon inférieure à 1kb, l'absence de dimérisation, et une température d'hybridation comprise entre 50 et 60 degrès (Tm). De plus, afin d'amplifier les différentes copies des gènes non synténiques (génome hexaploide), les amorces ont été définies manuellement, sous le programme ARTEMIS (Figure 16) dans des régions identiques à chaque génome (A/B/D).

b) Extraction d'ADN génomique

Selon le protocole d'extraction d'ADN de plante en grande quantité (MO-MOL-092), ADN de cinq grammes de feuilles de la lignée Chinese Spring ont été extrait. La quantité d'ADN extraite est mesurée par spectrophotométrie (Nanodrop®). De plus une première évaluation de la qualité de l'extraction est donnée par le ratio $D.O._{260}/D.O._{280}$. Ensuite la qualité des ADN est déterminée par électrophorèse sur gel d'agarose à 1%.

c) Test d'amplification PCR

Afin de valider les 53 couples d'amorces, une amplification par PCR de l'ADN génomique de Chinese Spring a été réalisée. Le milieu réactionnel de la PCR est composé de l'AmpliTaq Gold® 360 Master Mix, incluant la Taq polymerase et les désoxyribonucléotides, de 25ng d'ADN, de 5pmol et de 20pmol de SYTO9®. Le SYTO9® est un agent intercalant qui émet une fluorescence lorsqu'il est excité par une longueur d'onde de 500nm. Des cycles de température d'hybridation dégressifs ont été utilisés afin de permettre l'appariement de chaque couple d'amorces avec sa cible (PCR par essais).

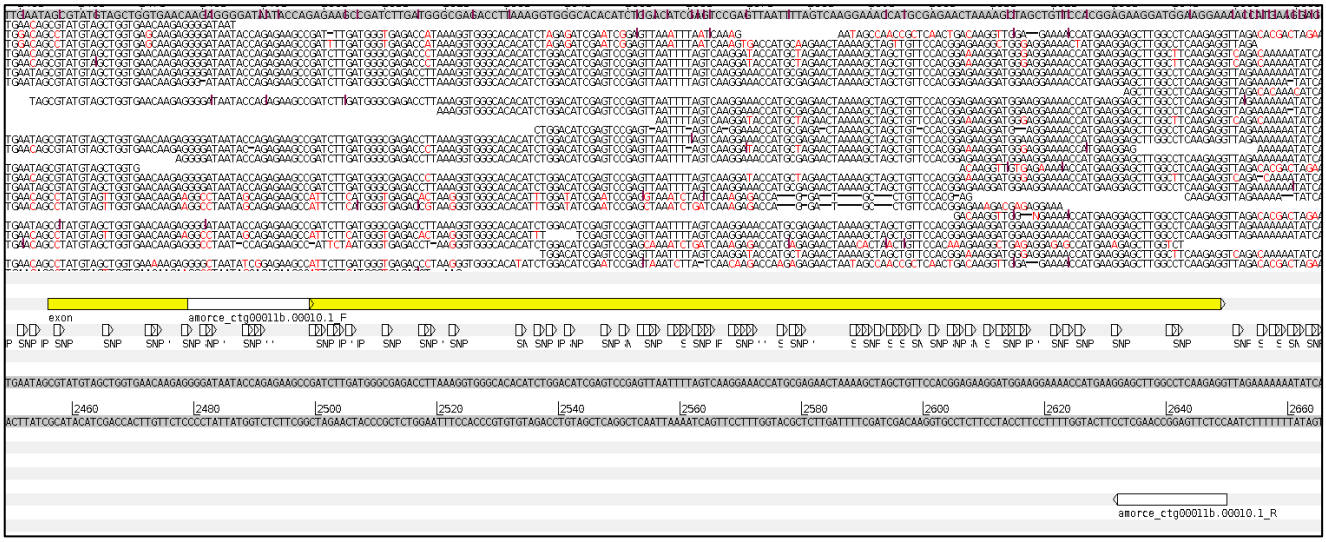


Figure 16 : Vue partielle d'un alignement de séquences 454 du set CS5x sur un gène de référence visualisé sous Artemis.

Un exon du gène ctg0011b.0010.1 est matérialisé par un rectangle jaune. La séquence de référence est indiquée en dessous (sur les 2 brins). Les séquences 454 du set CS5x alignées par Mosaik sont représentées en haut de la figure. Les substitutions entre les séquences 454 et le gène de référence apparaissent en rouge sur l'alignement et sous forme de rectangles blanc (SNP). Les polymorphismes SNP ont été identifiés par le programme GigaBayes.

Un couple d'amorces (rectangles blancs) a été dessiné dans les régions les plus conservées afin d'amplifier cet exon par PCR pour toutes les copies de ce gènes.

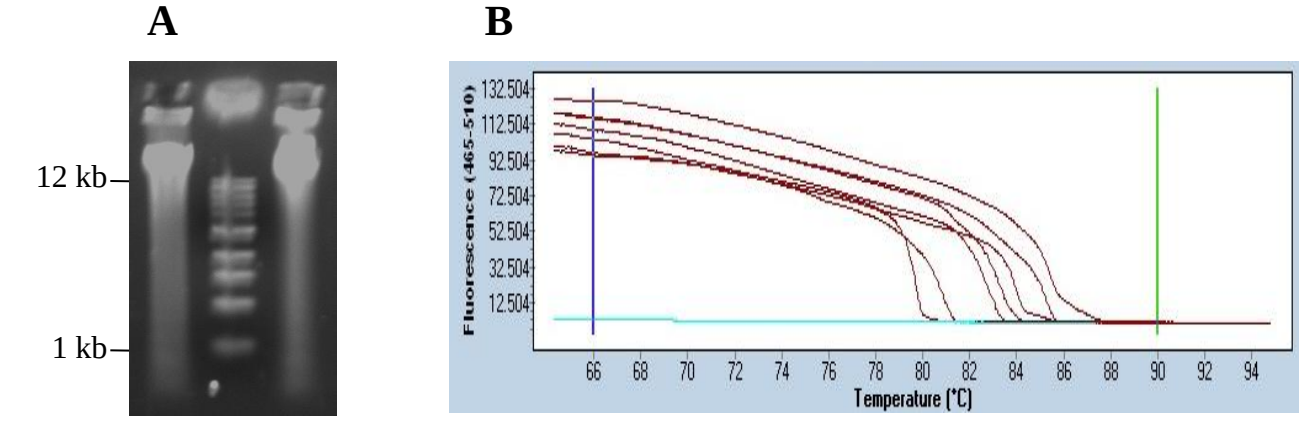


Figure 17 : Extraction d'ADN génomique et amplification par PCR.

(A) Electrophorèse sur gel d'agarose de l'ADN génomique extrait de la lignée Chinese Spring. Piste 2 : Marqueurs de taille, Piste 1 et 3 : ADN génomique.

(B) Exemples de courbes de dissociation obtenues après amplification de 7 produit de PCR sur l'ADN génomique de la variété Chinese Spring. Le témoin négatif (eau) ne présentant pas d'amplification est représenté par la courbe bleue.

d) Lecture des résultats

La lecture des courbes de dissociations est effectuée par un Light Cycler® 480 Real-Times PCR system (Roche applied science). Le principe de cette lecture est d'augmenter progressivement la température du milieu pour provoquer une dénaturation progressive des ADN doubles brins. Parallèlement la fluorescence émise par SYTO9® est observée à chaque seconde. Ainsi une courbe de dissociation des deux brins des fragments d'ADN peut être observée uniquement si le couple d'amorce s'est effectivement hybridé sur l'ADN génomique et que la réaction d'amplification s'est produite. Pour les 53 couples testés, une courbe de dissociation a été observée pour 51 couples.

DISCUSSION

1. Évaluation de la performance de l'approche mise au point et significativité des résultats obtenus

Au cours de cette étude, deux stratégies bioinformatiques basées sur l'alignement de séquences"shotgun" sur des séquences de gènes de référence ont été utilisées pour estimer le nombre de copies de 53 gènes potentiellement dupliqués. Les paramètres de recherche de similarité peuvent influencer significativement les estimations. Pour n'identifier que les copies récemment dupliquées, deux critères de sélection des séquences 454 ont été choisis : l'identité de séquences nucléotidiques et la longueur de l'alignement. Le pourcentage d'identité est un critère déterminant afin de n'identifier que les copies de gènes ayant été dupliquées récemment, donc n'ayant pas trop divergées. Nous avons testé les seuils 80 et 90% d'identité nucléotidique. Par ailleurs, la longueur minimale de l'alignement considérée comme significative influence grandement les résultats. En effet, seule la séquence codante (les exons) a été utilisée pour la recherche de similarité avec les séquences génomiques. Ainsi, la plupart des séquences 454 ne s'alignent que partiellement sur la CDS cible. Il a donc fallu intégrer les alignements partiels. Deux valeurs ont été testées : une longueur d'alignement minimale de 50 nucléotides ou supérieure à 40% de la taille de la séquence.

Parmi les quatre critères testés, le seuil de longueur fixé à 40% de la taille d'une séquence 454 est apparu comme non pertinent. En effet, les séquences 454 mesurent en moyenne 400 bp, ainsi, seules les séquences s'alignant sur au moins 160 pb (40%) sont sélectionnées. De grandes variations de la couverture ont été observées, notamment, pour les gènes composés de petits exons d'une taille inférieure à 160 bp (cas pour 86 gènes parmi les 134 étudiés). Avec ce critère, aucune similarité dans le set CS5x ne pouvait être retrouvée pour ces petits exons, démontrant un biais d'analyse. En utilisant une longueur seuil de 50 nucléotides, les taux de couverture observés étaient nettement plus homogènes. Il est donc apparu que ce critère était plus approprié ; toutefois, une région de faible complexité (microsatellite) présente dans un exon a posé problème et a dû être traitée manuellement.

Le choix du paramètre ne sélectionnant les séquences 454 que selon un certain niveau de similarité avec la séquence de référence a été largement influencé par le choix d'un seuil de longueur d'alignement supérieur à 50 nucléotides. En effet, sur une taille aussi petite, filtrer les lectures avec un seuil de 80% d'identité seulement (soit jusqu'à 10 substitutions/50 nucléotides) aurait conduit à augmenter le nombre d'alignements non spécifiques. Un seuil de 90% d'identité

a permis de contourner ce problème et de focaliser l'analyse sur les gènes fortement conservés, correspondant donc aux duplications les plus récentes.

Les deux approches utilisées (avec ou sans "clustering") pour estimer un nombre de copies de gènes présentent chacune des avantages et des inconvénients. Sans étape de "clustering", l'avantage est de pouvoir tenir compte des éventuelles copies de gènes ayant très peu divergé (duplication très récentes), mais l'inconvénient est que l'estimation est biaisée par les variations locales du taux de couverture. En effet, chaque région du génome n'est pas séquencée exactement 5 fois dans le set CS5x. Au contraire, l'étape de "clustering" permet de s'affranchir de ce problème, mais collapse les copies de gènes identiques. Ainsi, au vu de la variation de la couverture moyenne évaluée à 6+/-4x, l'approche par "clustering" apparaît comme la meilleure pour estimer le nombre de copies de gènes dupliqués dans le génome du blé. Toutefois, les différences dans l'estimation du nombre de copies entre les deux approches semblent être accentuées pour les gènes très fortement dupliqués. Quant aux gènes présents en faible nombre de copies, la concordance entre les deux estimations démontre la fiabilité des résultats.

2. Le génome des *Triticeae* est soumis à une intense activité de duplications de gènes

Les duplications de gènes constituent le moteur principal de la création de nouveaux gènes par mécanisme de duplication-divergence et, donc, de l'adaptation des organismes à leur environnement. La comparaison des premières grandes séquences contiguës du génome de blé (Choulet et al. 2010) avec les génomes séquencés du riz et de *Brachypodium*, a permis l'identification d'un set de 53 gènes non synténiques. Ces gènes récemment insérés à un nouveau locus sont répartis le long d'un squelette de gènes conservés de façon ancestrale, indiquant que chaque gène non synténique est issu d'un événement indépendant de duplication inter-chromosomique ayant été fixé spécifiquement dans la lignée des *Triticeae* après leur divergence avec *Brachypodium*. L'estimation du nombre de copies de gènes présents au sein du génome de blé a permis l'identification de trois pics avec un nombre de copies similaires.

Le premier groupe de gènes compte 2 à 4 copies. Ce nombre semble indiquer qu'un unique événement de duplication s'est produit spécifiquement après la séparation des 3 génomes A, B et D (soit chez le diploïde ancêtre du génome B, soit chez le polyploïde).

D'autre part, il est surprenant d'identifier plus de 50% des gènes appartenant à cette catégorie avec seulement 3 voire 2 copies. Mise à part une évaluation pessimiste du nombre de copie, ce résultat irait dans le sens d'un retour à l'état de trois copies homéologues voire à l'état dïploide se

manifestant par la perte de copies redondantes La duplication de certaines fonctions moléculaires peut, en effet, être contre sélectionnée.

Le deuxième groupe de gènes identifiés est présent en un nombre de copies allant de 8 à 9. Deux événements de duplications s'étant produits avant la divergence A-B-D, sans perte de copies homéologues, peuvent expliquer une telle redondance. Par conséquent, il apparaît intéressant d'approfondir nos connaissances sur ces gènes qui semblent avoir été fixés en un grand nombre de copies. Ils constituent un élément important du puzzle de la spécificité du blé vis-à-vis des autres *Poaceae*.

La dernière catégorie identifiée regroupe les gènes présents en un nombre de copies supérieur à 18. Cette catégorie représente 13% des gènes non synténiques. Un maximum de 122 copies est atteint pour un gène codant une histone H3, qui est, certes, connu pour être massivement dupliqué chez les *Poaceae*, et même au-delà. Les génomes du riz et de *Brachypodium* portent respectivement 14 et 13 copies (>80% d'identité nucléotidiques entre copies) de ce gène d'histone H3. Avec une valeur aussi élevée chez le blé, il paraît difficile d'affirmer avec certitude que ce nombre de copies est exact. Toutefois, il reflète une explosion du nombre de copies de gènes d'histones H3 dans la branche des *Triticeae*. Une valeur aussi élevée pourrait être expliquée par la mobilisation d'une copie par un élément transposable. Des analyses de séquences supplémentaires pourraient être réalisées afin de vérifier cette hypothèse. Par ailleurs, beaucoup de ces copies pourraient être inactives (pseudogènes). Des analyses du respect du cadre de lecture, ou des comparaisons du taux de substitutions synonymes (Ks) et non synonymes (Ka) pourraient également permettre d'identifier quelles copies ne sont plus soumises à pression de sélection.

Les deux mécanismes décrits comme principaux responsables des mouvements des gènes sont la capture d'ADN par les ET et la réparation de cassure double brin (Wicker et al. 2010). L'estimation du nombre de copies de gènes apparaît comme un bon indice pour tenter de détecter les mécanismes responsables des duplications inter-chromosomiques des 53 gènes non synténiques. Ainsi, il a été montré que la réparation de l'ADN double brin engendrerait un faible nombre de copies du gène dupliqué (Wicker et al. 2010). En effet, la réparation d'une cassure double brin est un événement unique, et la probabilité pour qu'un même fragment d'ADN contenant un gène soit dupliqué plusieurs fois est faible. Par conséquent, il semble probable que ce mécanisme ait été utilisé lors de la duplication des gènes non synténiques retrouvés en faible nombre de copies.

À l'inverse, le nombre de duplications générées par la capture de gènes par un ET est important. En effet, un gène capturé par un ET sera retrouvé en autant de copies que le nombre de transpositions de l'ET sur le génome. Ceci suggère que les gènes non synténiques retrouvés en un nombre de copie élevé aient été dupliqués par ce second mécanisme. De plus, parmi les 53 gènes non synténiques étudiés, deux sont intégrés dans des transposons CACTA : ctg0661b.00020, et ctg0954b.00040. Les estimations de leur nombre de copies sont de, respectivement, 19 et 20, confirmant une amplification massive des gènes mobilisés par des transposons.

Au total, 86% des gènes non synténiques pourraient être issus de la réparation d'une cassure double brin alors que seulement 14% seraient dûs à une capture par un élément transposable. Cela confirmerait l'hypothèse émise précédemment (Wicker et al. 2010) selon laquelle la capture de gènes par des TE serait moins fréquemment impliquée dans les duplications de gènes que les mécanismes de réparation de cassures double brin.

3. Perspectives

Afin d'identifier le locus ayant été dupliqué, une expérience basée sur l'amplification PCR des 53 gènes candidats sur les 42 bras de chromosomes triés en utilisant l'appareil de PCR micro-fluidique Biomark (Fluidigm Corporation).. Des amorces ont été dessinées dans les régions conservées des différentes copies. Cette expérience n'a pas donné de résultat concluant pour l'instant. Une mise au point de l'appareil est nécessaire. Elle est en cours au moment de l'écriture de ce rapport. Les résultats permettront, d'une part, de confirmer *in vitro* les estimations du nombre de copies de gènes, et, d'autre part, d'identifier le chromosome ancestral de chaque gène non synténique. Par la suite, un criblage des banques BAC disponibles pour chacun des bras de chromosome pourra permettre d'identifier des BAC porteurs de certaines copies dupliquées. Leur séquençage permettra une analyse évolutive plus précise des régions dupliquées.

La conservation de plusieurs copies de gènes redondants suggère que la subfonctionalisation jouerait un rôle majeur dans le maintien de l'intégrité structurale des gènes dupliqués. Des analyses fines de cinétique d'expression des gènes dupliqués seront entreprises afin de mieux caractériser la fonctionnalité des différentes copies. Les alignements réalisés au cours de cette étude permettront de dessiner des amorces spécifiques de chaque copie, ainsi utilisables pour tester l'expression de chaque copie individuellement par Q-PCR. Des données d'expression précises pour ces gènes non synténiques pourront permettre de vérifier l'hypothèse selon laquelle ces duplications auraient été sélectionnées .

BIBLIOGRAPHIE

2005. The map-based sequence of the rice genome. Nature. 436:793-800.

2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 463:763-8.

Abrouk, M., F. Murat, C. Pont, J. Messing, S. Jackson, T. Faraut, E. Tannier, C. Plomion, R. Cooke, C. Feuillet and J. Salse. 2010. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends Plant Sci. 15:479-87.

Akhunov, E.D., A.R. Akhunova, A.M. Linkiewicz, J. Dubcovsky, D. Hummel, G. Lazo, S. Chao, O.D. Anderson, J. David, L. Qi, B. Echalier, B.S. Gill, Miftahudin, J.P. Gustafson, M. La Rota, M.E. Sorrells, D. Zhang, H.T. Nguyen, V. Kalavacharla, K. Hossain, S.F. Kianian, J. Peng, N.L. Lapitan, E.J. Wennerlind, V. Nduati, J.A. Anderson, D. Sidhu, K.S. Gill, P.E. McGuire, C.O. Qualset and J. Dvorak. 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc Natl Acad Sci U S A. 100:10836-41.

Barnett, D.W., E.K. Garrison, A.R. Quinlan, M.P. Stromberg and G.T. Marth. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 27:1691-2.

Bateman, A., E. Birney, R. Durbin, S.R. Eddy, R.D. Finn and E.L. Sonnhammer. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. 27:260-2.

Baucom, R.S., J.C. Estill, J. Leebens-Mack and J.L. Bennetzen. 2009. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 19:243-54.

Bennett, M.D. and I.J. Leitch. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot. 107:467-590.

Bennetzen, J.L. 2007. Patterns in grass genome evolution. Curr Opin Plant Biol. 10:176-81.

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

Carver, T., M. Berriman, A. Tivey, C. Patel, U. Bohme, B.G. Barrell, J. Parkhill and M.A. Rajandream. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 24:2672-6.

Choulet, F., T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M.C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J.A. Anderson, B.S. Gill, R. Appels, B. Keller and C. Feuillet. 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell. 22:1686-701.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nat Rev Genet. 6:836-46.

Devos, K.M. 2010. Grass genome organization and evolution. Curr Opin Plant Biol. 13:139-45.

Devos, K.M., J.K. Brown and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12:1075-9.

Feschotte, C. and E.J. Pritham. 2009. A cornucopia of Helitrons shapes the maize genome. Proc Natl Acad Sci U S A. 106:19747-8.

Finn, R.D., J. Clements and S.R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res

Flavell, R.B., M.D. Bennett, J.B. Smith and D.B. Smith. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet. 12:257-69.

Gale, M.D. and K.M. Devos. 1998. Comparative genetics in the grasses. Proc Natl Acad Sci U S A. 95:1971-4.

Gordon, D., C. Abajian and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195-202.

Gregory, T.R., J.A. Nicol, H. Tamm, B. Kullman, K. Kullman, I.J. Leitch, B.G. Murray, D.F. Kapraun, J. Greilhuber and M.D. Bennett. 2007. Eukaryotic genome size databases. Nucleic Acids Res. 35:D332-8.

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

Hufton, A.L. and G. Panopoulou. 2009. Polyploidy and genome restructuring: a variety of outcomes. Curr Opin Genet Dev. 19:600-6.

Jiang, N., Z. Bao, X. Zhang, S.R. Eddy and S.R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 431:569-73.

Jiang, Z., R. Hubley, A. Smit and E.E. Eichler. 2008. DupMasker: a tool for annotating primate segmental duplications. Genome Res. 18:1362-8.

Kapitonov, V.V. and J. Jurka. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. 23:521-9.

Keller, B. and C. Feuillet. 2000. Colinearity and gene density in grass genomes. Trends Plant Sci. 5:246-51.

Lai, J., Y. Li, J. Messing and H.K. Dooner. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci U S A. 102:9068-73.

Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 2:231-9.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078-9.

Liu, R., C. Vitte, J. Ma, A.A. Mahama, T. Dhliwayo, M. Lee and J.L. Bennetzen. 2007. A GeneTrek analysis of the maize genome. Proc Natl Acad Sci U S A. 104:11844-9.

Mayer, K.F., M. Martis, P.E. Hedley, H. Simkova, H. Liu, J.A. Morris, B. Steuernagel, S. Taudien, S. Roessner, H. Gundlach, M. Kubalakova, P. Suchankova, F. Murat, M. Felder, T. Nussbaumer, A. Graner, J. Salse, T. Endo, H. Sakai, T. Tanaka, T. Itoh, K. Sato, M. Platzer, T. Matsumoto, U. Scholz, J. Dolezel, R. Waugh and N. Stein. 2011. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell. 23:1249-63.

Mayer, K.F., S. Taudien, M. Martis, H. Simkova, P. Suchankova, H. Gundlach, T. Wicker, A. Petzold, M. Felder, B. Steuernagel, U. Scholz, A. Graner, M. Platzer, J. Dolezel and N. Stein. 2009. Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 151:496-505.

Paterson, A.H., J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A.K. Bharti, J. Chapman, F.A. Feltus, U. Gowik, I.V. Grigoriev, E. Lyons, C.A. Maher, M. Martis, A. Narechania, R.P. Otillar, B.W. Penning, A.A. Salamov, Y. Wang, L. Zhang, N.C. Carpita, M. Freeling, A.R. Gingle, C.T. Hash, B. Keller, P. Klein, S. Kresovich, M.C. McCann, R. Ming, D.G. Peterson, R. Mehboob ur, D. Ware, P. Westhoff, K.F. Mayer, J. Messing and D.S. Rokhsar. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature. 457:551-6.

Paterson, A.H., J.E. Bowers, B.A. Chapman, D.G. Peterson, J. Rong and T.M. Wicker. 2004. Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. Curr Opin Biotechnol. 15:120-5.

Paux, E., D. Roger, E. Badaeva, G. Gay, M. Bernard, P. Sourdille and C. Feuillet. 2006. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J. 48:463-74.

Paux, E., P. Sourdille, J. Salse, C. Saintenac, F. Choulet, P. Leroy, A. Korol, M. Michalak, S. Kianian, W. Spielmeyer, E. Lagudah, D. Somers, A. Kilian, M. Alaux, S. Vautrin, H. Berges, K. Eversole, R. Appels, J. Safar, H. Simkova, J. Dolezel, M. Bernard and C. Feuillet. 2008. A physical map of the 1-gigabase bread wheat chromosome 3B. Science. 322:101-4.

Petrov, D.A. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. 17:23-8.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, H. Kim, K. Collura, D.S. Brar, S. Jackson, R.A. Wing and O. Panaud. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 16:1262-9.

Prasad, V., C.A. Stromberg, H. Alimohammadian and A. Sahni. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science. 310:1177-80.

Qi, L.L., B. Echalier, S. Chao, G.R. Lazo, G.E. Butler, O.D. Anderson, E.D. Akhunov, J. Dvorak, A.M. Linkiewicz, A. Ratnasiri, J. Dubcovsky, C.E. Bermudez-Kandianis, R.A. Greene, R. Kantety, C.M. La Rota, J.D. Munkvold, S.F. Sorrells, M.E. Sorrells, M. Dilbirligi, D. Sidhu, M. Erayman, H.S. Randhawa, D. Sandhu, S.N. Bondareva, K.S. Gill, A.A. Mahmoud, X.F. Ma, Miftahudin, J.P. Gustafson, E.J. Conley, V. Nduati, J.L. Gonzalez-Hernandez, J.A. Anderson, J.H. Peng, N.L. Lapitan, K.G. Hossain, V. Kalavacharla, S.F. Kianian, M.S. Pathan, D.S. Zhang, H.T. Nguyen, D.W. Choi, R.D. Fenton, T.J. Close, P.E. McGuire, C.O. Qualset and B.S. Gill. 2004. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics. 168:701-12.

Rustenholz, C., P.E. Hedley, J. Morris, F. Choulet, C. Feuillet, R. Waugh and E. Paux. 2010. Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources. BMC Genomics. 11:714.

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

SanMiguel, P.J., W. Ramakrishna, J.L. Bennetzen, C.S. Busso and J. Dubcovsky. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct Integr Genomics. 2:70-80.

Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T.A. Graves, P. Minx, A.D. Reily, L. Courtney, S.S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S.M. Rock, E. Belter, F. Du, K. Kim, R.M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S.M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M.J. Levy, L. McMahan, P. Van Buren,

M.W. Vaughn, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112-5.

Stajich, J.E. 2007. An Introduction to BioPerl. Methods Mol Biol. 406:535-48.

Tenaillon, M.I., J.D. Hollister and B.S. Gaut. 2010. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 15:471-8.

Vitte, C. and J.L. Bennetzen. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci U S A. 103:17638-43.

Wicker, T., J.P. Buchmann and B. Keller. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. 20:1229-37.

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

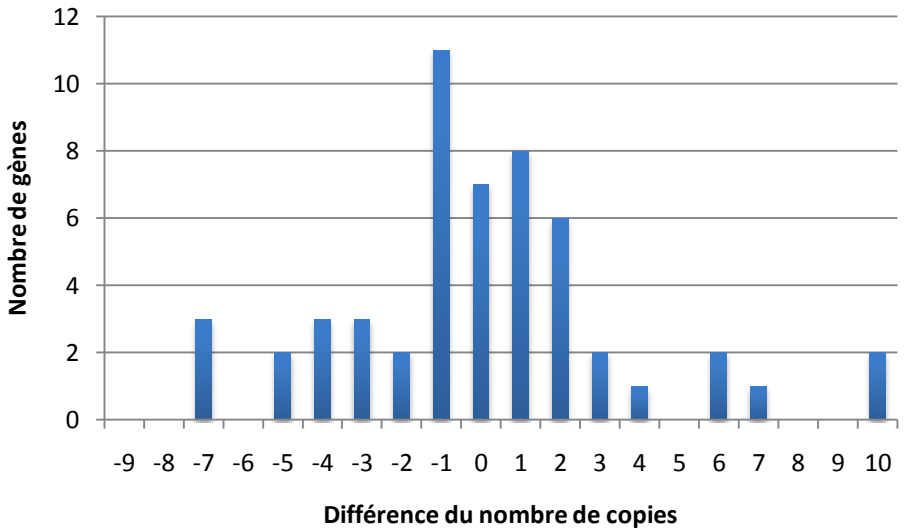
Wicker, T., K.F. Mayer, H. Gundlach, M. Martis, B. Steuernagel, U. Scholz, H. Simkova, M. Kubalakova, F. Choulet, S. Taudien, M. Platzer, C. Feuillet, T. Fahima, H. Budak, J. Dolezel, B. Keller and N. Stein. 2011. Frequent Gene Movement and Pseudogene Evolution Is Common to the Large and Complex Genomes of Wheat, Barley, and Their Relatives. Plant Cell

Yang, L. and J.L. Bennetzen. 2009. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A. 106:19922-7.

Yu, Z., S.I. Wright and T.E. Bureau. 2000. Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics. 156:2019-31.

Chromosome du blé	Nombre de gènes
1	9
2	6
3	-
4	14
5	10
6	5
7	9

Annexe 1 : Localisation chromosomique prédite de la copie ancestrale des 53 gènes non synténiques ayant été dupliquée.
La localisation chromosomique a été déduite de la synténie entre les génomes du blé et du riz précédemment mise en évidence à l'échelle des génomes complets. Pour chaque gène non synténique identifié sur le chromosome 3B, le meilleur homologue identifié chez le riz a été considéré pour prédire la localisation du locus ayant été dupliqué.



Annexe 2 : Distribution des différences du nombre de copies de gènes non synténiques estimées avant et après "clustering".
Une valeur de zéro indique que les 2 estimations sont identiques. Les valeurs positives indiquent un nombre de copies plus élevé avant clustering et inversement.

Gènes	Nombre de copies	Fonctions
ctg0011b.00010	11	male sterility protein, putative, expressed
ctg0011b.00040	8	male sterility protein, putative, expressed
ctg0011b.00060	9	Flowering Locus T-like protein, putative
ctg0011b.00070	15	Flowering Locus T-like protein, putative
ctg0011b.00100	4	PPR repeat domain containing protein
ctg0011b.00110	15	Flowering Locus T-like protein, putative
ctg0011b.00120	4	PPR repeat domain containing protein
ctg0011b.00140	6	Tetratricopeptide repeat domain containing protein, expressed
ctg0011b.00180	4	fasciclin-like arabinogalactan precursor, putative, expressed
ctg0011b.00190	6	beta-expansin 1a precursor, putative, expressed
ctg0011b.00200	8	beta-expansin 1a precursor, putative, expressed
ctg0011b.00210	10	disulfide oxidoreductase, putative, expressed
ctg0011b.00230	3	Targeting protein for Xklp2 domain containing protein, expressed
ctg0011b.00250	8	hydroquinone glucosyltransferase, putative, expressed
ctg0011b.00340	8	Paf1 domain containing protein, expressed
ctg0079b.00010	5	sedoheptulose-1,7-bisphosphatase, chloroplast precursor, expressed
ctg0091b.00060	4	hypothetical protein
ctg0091b.00080	6	acyl-desaturase, chloroplast precursor, putative, expressed
ctg0091b.00150	9	ubiquitin-protein ligase, putative, expressed
ctg0091b.00160	9	ubiquitin-protein ligase, putative, expressed
ctg0382b.00140	5	F-box domain containing protein, expressed
ctg0464b.00020	122	histone H3, expressed
ctg0464b.00070	4	Tuftelin interacting protein 11 domain containing protein
ctg0464b.00080	14	F-box domain containing protein, expressed
ctg0464b.00160	9	F-box domain containing protein
ctg0464b.00180	24	peroxidase 66 precursor, putative, expressed
ctg0464b.00210	4	F-box domain containing protein, expressed
ctg0616b.00200	3	50S ribosomal protein L15, putative, expressed
ctg0661b.00020	19	cytochrome P450, putative, expressed
ctg0661b.00030	9	ABC transporter domain containing protein, expressed
ctg0661b.00040	32	beta-expansin 1a precursor, putative, expressed
ctg0661b.00050	18	cytochrome P450, putative, expressed
ctg0954b.00010	5	xyloglucan galactosyltransferase KATAMARI 1, putative, expressed
ctg0954b.00040	20	conserved hypothetical protein, expressed
ctg0954b.00050	3	glutathione S-transferase, putative, expressed
ctg0954b.00070	4	conserved hypothetical protein, expressed
ctg0954b.00090	3	receptor-like protein kinase, putative, expressed
ctg0954b.00130	6	protein kinase, putative, expressed
ctg0954b.00170	8	conserved hypothetical protein, expressed
ctg0954b.00180	3	conserved hypothetical protein, expressed
ctg0954b.00200	12	glutamate decarboxylase, putative, expressed
ctg0954b.00215	8	conserved hypothetical protein, Hv-pg1 homolog, putative, expressed
ctg0954b.00220	7	conserved hypothetical protein, Hv-pg4 homolog, expressed
ctg0954b.00260	9	sodium/hydrogen exchanger, putative, expressed
ctg0954b.00270	29	anaphase-promoting complex subunit, putative, expressed
ctg0954b.00280	3	glycosyltransferase, HGA-like, putative, expressed
ctg0954b.00300	8	exonuclease, putative
ctg0954b.00340	11	tRNA (guanine-N(1)-)-methyltransferase, putative, expressed
ctg0954b.00360	10	polygalacturonase precursor, putative, expressed
ctg0954b.00380	3	terpene synthase, putative
ctg0954b.00460	10	cytochrome P450, putative
ctg0954b.00470	3	NB-ARC domain containing protein
ctg0954b.00480	2	F-box domain containing protein

Annexe 3 : Tableau répertoriant le nombre de copies des 53 gènes non synténique ainsi que leurs fonctions

Programmes publics	Fonctions	Options
fastacmd	Extrait les séquences choisies d'une base de donnée	-i liste des numéros d'identité des séquences à extraire
RepeatMasker	Transforme les nucléotides des éléments transposables identifiés en minuscule dans le but qu'elles soient masquées et non prises en compte dans les analyses futures.	-d base de donnée répertoriant tout les familles d'éléments transposables identifiés.
ARTEMIS	Navigateur permettant la visualisation d'un fichier d'annotation embl, fasta, gff, et bam.	
BLAST	Rechercher les similarités entre une séquence fournie et toutes les séquences d'une base de données	-p : type de blast -i : séquence de référence -d base de donnée -e : score au dessus duquel les alignements ne sont pas reportés
Mosaik	Aligne des séquences nucléotidiques sur une séquence de référence	-mmp : spécifie le pourcentage maximum de la longueur de la séquence à aligner qui peut être différent de la séquence de référence -mmal : tolère les alignements débordant la séquence de référence
Gigabayse	Programme permettant la détection de SNP	--CRL : couverture minimale de la séquence de référence pour pouvoir détecter une SNP --CAL : nombre de fois qu'un SNP est observé pour être identifié
Phrap	Programme permettant la création de contigs sur les similarités de chevauchement des séquences.	

Annexe 4 : Tableau référençant les principaux programmes publics utilisés dans les analyses

Programmes développés dans l'unité	Fonctions	Options
embl2fna.pl	Extrait les séquences d'un fichier embl pour les convertir en un format fasta	-i : Liste des numéros d'identité des séquences a extraire
ParseBlast.pl	Analyse un fichier créé par le programme BLAST afin de trier les alignements sur certains critères	-i : identité -qov : pourcentage de la longueur de la séquence de référence s'alignant sur la séquence alignée -l : nombres de nucléotides incluses dans l'alignement pour qu'il soit reporté
BlastMap	Pipeline d'alignement de séquences	-conf : fichier de configuration
calc_coverage.pl	Calcule la couverture d'une séquence de référence à partir d'un fichier créé par MosaikCoverage	-gff : position des exons de la séquence de référence

Annexe 5 : Tableau référençant les principaux programmes développés dans l'unité utilisés dans les analyses

Annexe 6 : Code perl du pipeline BlastMap (A) et du fichier de configuration en xml (B)

(A) Code perl du pipeline BlastMap.

```
#!/usr/bin/perl
use strict;
use warnings;
use File::Basename;
use Getopt::Long;
use Data::Dumper;
use XML::Twig ;
use Bio::SeqIO ;
my $VERSION = '1.0' ;
my $lastmodif = '2011-4-22' ;
my $help ;
my $conf ;
my $db ;
my $ref_opt ;
my $blastoutfile ;
my $fastareads ;

&GetOptions(
      "h"      => \$help,
      "conf=s"   => \$conf,
      "db=s"         => \$db,
      "ref=s"    => \$ref_opt,
      "blast=s"  => \$blastoutfile,
      "reads=s"   => \$fastareads
);

$help and &help ;
@ARGV or &help ;

-e $conf or die "*** could not find config file: \"$conf\" ***\n" ;
unless (defined $fastareads and -e $fastareads){$db or die "*** Blast database
\"$db\" not found\n" ;}

################################

###Parsing configuration file ###

################################
my %hash_param ;
my $xml = $conf ;
my $twig = XML::Twig->new;
$twig->parsefile( $xml );
my $config = $twig->first_elt( 'config' );
foreach my $program_obj ($config->children('program')) {
      my $program_name = $program_obj->att('name') ;
      foreach my $param_obj ($program_obj->children('parameter')) {
            my $name  = $param_obj->att('name') ;
            my $value = $param_obj->att('value') ;
            $hash_param{$program_name}{$name} = $value ;
      }
}
################################
### Main analysis             ###
################################
my %HashRefSeqObj ;
if (defined $ref_opt and -e $ref_opt) {
      my $SeqIoObj = Bio::SeqIO->new('-file'=>$ref_opt, '-format'=>"FASTA") ;
```

```
	while(my $SeqObj = $SeqIoObj->next_seq() ) {
		$HashRefSeqObj{$SeqObj->display_id()} = $SeqObj ;
	}
}
foreach my $SeqToBlast (@ARGV) {
	-e $SeqToBlast or die "*** file \"$SeqToBlast\" not found ***\n";

	my $SeqIoObj = Bio::SeqIO->new('-file'=>$SeqToBlast, '-format'=>"FASTA") ;
	while(my $SeqObj = $SeqIoObj->next_seq() ) {
		my $id = $SeqObj->display_id() ;
		my $PrefixFileName = basename($id) ;
		my $SingleFastaFileCDS = $id.'.blast.fasta' ;
		my $outSeqIoObj = Bio::SeqIO->new('-file'=">$SingleFastaFileCDS", '-
format'=>"FASTA") ;
		$outSeqIoObj ->write_seq($SeqObj) ;
		print STDERR "--> Create file $SingleFastaFileCDS\n" ;

#		my $PrefixFileName = basename($id) ;
#		$PrefixFileName =~s/\.[^\.]*$// ;
		unless (defined $fastareads and -e $fastareads){
			unless (-e $blastoutfile){
				##Blast##
				print "--> runing Blastall...\n";
				$blastoutfile =
&launchblastall($SingleFastaFileCDS,$db,$PrefixFileName) ;
#				unlink $SingleFastaFileCDS ;
			}
			##Parseblast##
			print "--> runing ParseBlast.pl...\n";
			my $IdBlast = &launchparseblast($blastoutfile,$id) ;

			##fastacmd##
			print"--> runing fastacmd...\n";
			$fastareads = &launchfastacmd($IdBlast,$db,$PrefixFileName) ;
			unlink $IdBlast ;
		}

		my $refseq = $SingleFastaFileCDS ;
		my $SingleFastaFileGene = $id.".gene.fasta" ;
		if ($ref_opt) {
			if ($HashRefSeqObj{$id}) {
				my $outRefSeqIoObj = Bio::SeqIO->new('-
file'=>">$SingleFastaFileGene", '-format'=>"FASTA") ;
				$outRefSeqIoObj ->write_seq ( $HashRefSeqObj{$id} ) ;
				$refseq = $SingleFastaFileGene ;
				print STDERR "--> Create file $SingleFastaFileGene\n" ;
			}
			else {
				print STDERR "*** sequence: \"$id\" not found in $ref_opt -
> skip ***\n" ;
				next ;
			}
		}
		##MOSAIK##
		#Sequence reference
		my $BuildrefOut = &launchMosaikBuild($refseq, "ref", $PrefixFileName) ;
#		unlink $SingleFastaFileGene if (-e $SingleFastaFileGene) ;
		#Reads

		my $BuildreadsOut = &launchMosaikBuild($fastareads, "reads",
$PrefixFileName);
```

```
            my $alignedOut = &launchMosaikAligner($BuildreadsOut, $BuildrefOut,
$PrefixFileName) ;
            print "created file $alignedOut \n" ;
#           unlink $BuildreadsOut ;

            my $sortedOut = &launchMosaikSort($alignedOut, $PrefixFileName) ;
            print "created file $sortedOut \n" ;
            unlink $alignedOut ;

            my $assemblyOut = &launchMosaikAssembler($sortedOut, $BuildrefOut,
$PrefixFileName);
            my $gigOut = "Assembly.gig_".basename($id).".gig" ;
            print "created file $gigOut \n" ;
#           unlink $BuildrefOut ;

            my $bamOut = &launchMosaikText($sortedOut, $PrefixFileName);
            print "created file $bamOut \n" ;

            &launchsamtools($bamOut);

            &launchMosaikCoverage ($sortedOut, $BuildrefOut, $PrefixFileName);

            ##gigaBayes## ($assemblyOut shoud be a gig file)
            &launchgigaBayes ($gigOut, $PrefixFileName) ;

            print STDERR "Seq: $id => done\n" ;
      } # END OF WHILE next_seq

      print STDERR "File: $SeqToBlast => done\n" ;
} # end of foreach SeqToBlast
###################
### Subroutines ###
###################
#---------------
# launchblastall
#---------------
sub launchblastall{
      my $inFile = $_[0] ;
      my $db = $_[1] ;
      my $blastoutFile = $_[2].".".$hash_param{"blastall"}{"p"} ;
      my @blastall = ('blastall', '-i' , $inFile , '-d' , $db, '-o' , $blastoutFile
) ;
      foreach (keys %{$hash_param{"blastall"}}) {
            push (@blastall, "-".$_) ;
            push (@blastall, $hash_param{"blastall"}{$_}) ;
      }
      system (@blastall) ;
      print join ( " ", @blastall),"\n" ;
      return $blastoutFile ;
}#end of sub launcherblastall
#---------------
# ParseBlast
#---------------
sub launchparseblast{
      my $inFile = $_[0] ;
      my $ParseBlastOutFile = basename($inFile).".parse" ;
      my $id = $_[1] ;
      -e $inFile or die "*** sub launchparseblast file \"$inFile\" doesn't exist
\n";
      my @parseblast = ("ParseBlast.pl") ;
      foreach (keys %{$hash_param{"ParseBlast.pl"}}) {
```

```
            push (@parseblast, "-".$_) ;
            push (@parseblast, $hash_param{"ParseBlast.pl"}{$_}) ;
      }
      push (@parseblast, ( $inFile, ">", $ParseBlastOutFile)) ;
      print join(" ", @parseblast), "\n" ;
      system(join" ", @parseblast) ;
      open (PARSEBLAST, "<$ParseBlastOutFile") or die "Can't open file
\"$ParseBlastOutFile\"\n";
      my $IdBlast = $ParseBlastOutFile.".lst" ;
      open (ID, ">$IdBlast") or die "Can't create file \"$IdBlast\"\n";
      print "--> Get Id of blast hits\n" ;
      while(<PARSEBLAST>) {
            chomp $_ ;
            if($_=~/^#/) { next ; }
            my @Parse = split (/\t/, $_) ;
            $Parse[0] eq $id or next ;
            $Parse[0] eq "no_hit" and next ;
            print ID "$Parse[1]\n" ;
      } # end of while
      close ID ;
      close PARSEBLAST ;
      return $IdBlast ;
}#end of sub launcherparseblast
#---------------
#fastacmd
#---------------
sub launchfastacmd{
      my $inFile = $_[0] ;
      my $db = $_[1] ;
      my $outFile = $_[2].".reads.fna" ;
      -e $inFile or die "*** sub launchfastacmd: infile \"$inFile\" doesn't exist
\n" ;
      my @fastacmd = ( 'fastacmd' ,
                       '-i' , $inFile,
                       '-d' , $db,
                       '-o' , $outFile ) ;
      print join(" ", @fastacmd), "\n" ;
      system (join(" ", @fastacmd)) ;
      return $outFile ;
}#end of sub launcherfastacmd
#---------------
#MosaikBuild
#---------------
sub launchMosaikBuild{
      print "--> runing MosaikBuild...\n";
      my $inFile = $_[0];
      my $seqtype = $_[1];
      my $outBuild = $_[2] . "." . $seqtype . ".dat" ;
      -e $inFile or die "*** sub launchMosaikBuild: infile \"$inFile\" doesn't
exist \n";
      my @mosaikBuild = ( 'MosaikBuild', '-fr', $inFile ) ;
      if ($seqtype eq 'ref') {
            push (@mosaikBuild, ('-oa' , $outBuild ,
                                 '-assignQual' ,
$hash_param{"MosaikBuild"}{"assignQual"},
                                 '>',
                                 'mosaikBuildref.log') ) ;
      }
      else {
            print "--> runing fasta2qual.pl...\n";
            my $readqual = &launchfasta2qual($inFile) ;
            push (@mosaikBuild, ('-out' , $outBuild ,
```

```
                                    '-st', $hash_param{'MosaikBuild'}{'st'},
                                    '-fq' , $readqual,
                                    '>',
                                    'mosaikBuildreads.log' ) ) ;
      }
      print join(" ", @mosaikBuild), "\n" ;
      system(join" ", @mosaikBuild) ;
      return $outBuild ;
} # end of sub launchMosaikBuild
#---------------
#MosaikAligner
#---------------
sub launchMosaikAligner{
      print "--> runing MosaikAligner...\n";
      my $inFile = $_[0] ;
      my $iaFile =  $_[1];
      my $outFile;
      $inFile or die "*** sub launchMosaikAligner: infile \"$inFile\" doesn't exist
\n" ;
      $outFile = $_[2].".Aligned.dat" ;
      my @mosaikAligner = ( 'MosaikAligner' , '-in' , $inFile , '-out' , $outFile ,
'-ia' , $iaFile) ;
      foreach (keys %{$hash_param{"MosaikAligner"}}) {
            push (@mosaikAligner, "-".$_) ;
            push (@mosaikAligner, $hash_param{"MosaikAligner"}{$_}) unless ($_ eq
"mmal");
      }
      push (@mosaikAligner, '>',
                           'mosaikAligner.log' ) ;
      system(join" ", @mosaikAligner) ;
      return $outFile ;
} # end of sub launchMosaikAligner
#---------------
#MosaikSort
#---------------
sub launchMosaikSort {
      print "--> runing MosaikSort...\n";
      my $inFile = $_[0]  ;
      my $outFile = $_[1].".Sorted.dat" ; ;
      -e $inFile or die "*** sub launchMosaikSort: infile \"$inFile\" doesn't exist
\n" ;
      my @mosaikSort = ( 'MosaikSort' , '-in' , $inFile , '-out' , $outFile) ;
      push (@mosaikSort, '>',
                          'mosaikSort.log' ) ;
      system(join" ", @mosaikSort) ;
      return $outFile ;
} # end of sub launchMosaikSort
#---------------
#MosaikAssembler
#---------------
sub launchMosaikAssembler {
      print "--> runing MosaikAssembler...\n";
      my $inFile =  $_[0] ;
      my $iaFile = $_[1] ;
      -e $inFile or die "*** sub launchMosaikAssembler: infile \"$inFile\" doesn't
exist \n" ;
      my $outFile = "Assembly" ;
      $outFile .= ".gig" if ($hash_param{"MosaikAssembler"}{"fileformat"} eq 'gig')
;
      $outFile .= ".ace" if ($hash_param{"MosaikAssembler"}{"fileformat"} eq 'ace')
;
      my @mosaikAssembler = ( 'MosaikAssembler' ,
```

```
                        '-in' , $inFile ,
                        '-ia' , $iaFile ,
                        '-out' , $outFile ,
                        '-f' , $hash_param{"MosaikAssembler"}{"fileformat"},
                              '>',
                              'mosaikAssembly.log' ) ;
      system(join" ", @mosaikAssembler) ;
      print join(" ", @mosaikAssembler), "\n" ;
      return ($outFile) ;
}# end of sub launchMosaikAssembler
#---------------
#MosaikText
#---------------
sub launchMosaikText {
      print "--> runing MosaikText...\n";
      my $inFile =  $_[0] ;
      my $bamFile ;
      -e $inFile or die "*** sub launchMosaikText: infile \"$inFile\" doesn't exist
\n";
            $bamFile = $_[1].".Sorted.bam" ;
            my @mosaikText = ( 'MosaikText' , '-in' , $inFile , '-bam' , $bamFile)
;
            push (@mosaikText, '>',
                        'mosaikText.log' ) ;
      system(join" ", @mosaikText) ;
      return $bamFile ;
}# end of sub launchMosaikText
#---------------
#samtools
#---------------
sub launchsamtools {
      my $inFile = $_[0] ;
      -e $inFile or die "*** sub launchsamtools: infile \"$inFile\" doesn't exist
\n";
            `samtools index $inFile` ;
} # end of sub launchsamtools
#---------------
#MosaikCoverage
#---------------
sub launchMosaikCoverage {
      print "--> runing MosaikCoverage...\n";
      my $inFile = $_[0] ;
      my $iaFile = $_[1] ;
      -e $inFile or die "*** sub launchMosaikCoverage: infile \"$inFile\" doesn't
exist \n";
      my @mosaikCoverage = ( 'MosaikCoverage' , '-in' , $inFile , '-ia' , $iaFile,
'-cg' ) ;
      push (@mosaikCoverage, '>',
                        'mosaikCoverage.log' ) ;
      system(join" ", @mosaikCoverage) ;
}# end of sub launchMosaikCoverage
#---------------
#gigaBayes
#---------------

sub launchgigaBayes{
      print "--> runing gigaBayes...\n";
      print "$_[0]\n" ;
      my $inFile = $_[0] ;
      my $outFile = $_[1].'.gff' ;
      -e $inFile or die "*** sub launchgigaBayes: infile \"$inFile\" doesn't exist
!!!\n" ;
```

```
      my @gigaBayes = ( 'gigaBayes', '--gig' , $inFile , '--gff' , $outFile) ;
      foreach (keys %{$hash_param{"gigaBayes"}}) {
            push (@gigaBayes, "--".$_) ;
            push (@gigaBayes, $hash_param{"gigaBayes"}{$_}) ;
      system (@gigaBayes) ;
            }
}
#---------------
#FASTA2QUAL
#---------------
sub launchfasta2qual{
      my $inFile = $_[0] ;
      -e $inFile or die "*** sub launchfasta2qual: infile \"$inFile\" doesn't exist
\n";
            my @fasta2qual ;
            @fasta2qual = ( 'fasta2qual.pl' , '-q' , 40, '-o', $inFile) ;
            system (@fasta2qual) ;
            return $inFile.".qual" ;

}# end of sub launchfasta2qual
#================================================================================
sub help {
my $prog = basename($0) ;
print STDERR <<EOF ;
#### $prog ####
#
# CREATED:    2011-4-19
# LAST MODIF: $lastmodif
# AUTHOR:     Josquin DARON (INRA Clermont-Ferrand)
# VERSION:    $VERSION
#
# This script is used aligne reads to a reference sequence
#Step : Blast, ParseBlast, Mosaik, gigaBayes
USAGE:
       $prog  [OPTIONS]  -db Database FASTA_file  ...

       ### OPTIONS ###
       -db string:     Database
       -ref string:    FASTA file containing reference sequences for read mapping
by Mosaik
                       FASTA files given as arguments are used for BLAST search
against -db
                       but sequences used for mapping matching reads can be
different.
                       ex: Blast with CDSs (exons only), Map reads on genes
(exons+introns)
       -blast string:  give blast file instead of runing blast
       -reads string:  give fasta reads file

EOF
exit(1) ;
}
#================================================================================
```

(B) Fichier de configuration au format xml.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<config>
      <program name="blastall" >
            <parameter name="e"            value="1e3" />  <!-- evalue -->
            <parameter name="p"          value="blastn" />
            <parameter name="F"          value="'m D'" />    <!-- filter -->
            <parameter name="v"          value="10000" />
            <parameter name="b"          value="10000" />
            <parameter name="U"          value="T" />
            <parameter name="A"          value="12" />
      </program>
      <program name="ParseBlast.pl" >
            <parameter name="o"          value="hsp" />
            <parameter name="i"          value="90" />
<!--        <parameter name="hov"        value="40" /> -->
            <parameter name="n"          value="100000000" />
            <parameter name="f"          value="tab" />
            <parameter name="l"          value="50" />
            <parameter name="1"          value="" />
      </program>
      <program name="fastacmd" >
      </program>
      <program name="MosaikBuild" >
            <parameter name="assignQual"  value="40" />
            <parameter name="st"         value="454" />
      </program>
      <program name="MosaikAligner" >
            <parameter name="hs"         value="15" />
      <parameter name="act"         value="20" />
            <parameter name="p"          value="2" />
<!--        <parameter name="mm"         value="4" /> -->
            <parameter name="mmp"        value="0.4" />
            <parameter name="mmal"       value="" />
            <parameter name="minp"       value="0.5" />
            <parameter name="mhp"        value="100" />
            <parameter name="m"          value="unique" />
      </program>
      <program name="MosaikAssembler" >
            <parameter name="fileformat"  value="gig" />
      </program>
      <program name="gigaBayes" >
            <parameter name="CRL"        value="5" />
            <parameter name="CAL"        value="1" />
      </program>
</config>
```

Résumé

Le séquençage de 18 Mb du chromosome 3B du blé tendre a permis de mettre en évidence qu'une proportion significative des gènes identifiés ne sont pas conservés en position synténique chez les génomes apparentés du riz, de *Brachypodium* et du sorgho. Ces gènes ont donc été insérés à un nouveau locus (sur le 3B) récemment, après la divergence du blé et de *Brachypodium*, il y a moins de 20 à 30 millions d'années. L'étude réalisée au cours de ce stage vise à caractériser l'origine de ces gènes et leur nombre de copies dans le génome hexaploïde du blé. Au total, 53 gènes non synténiques et 81 gènes synténiques ont été utilisés comme set de référence pour étudier l'ampleur des duplications de gènes sur l'évolution du génome du blé. Une approche bioinformatique a ainsi été mise au point. A partir des données de séquençage "shotgun" du génome de blé à un taux de couverture de 5x, le pipeline *BlastMap* a été développé pour estimer le nombre de copies dupliquées de chacun de ces gènes. Les résultats montrent que les gènes non synténiques sont plus massivement dupliqués au sein du génome que les gènes synténiques. Ainsi, 2/3 d'entre eux sont présents en 6 copies ou plus, et 13% sont présents en plus de 18 copies dans le génome, suggérant des mécanismes de duplication via la mobilisation par des éléments transposables. Cette étude révèle la fréquence très élevée des duplications inter-chromosomiques chez les *Triticeae* par rapport aux espèces apparentées.

Abstract

The sequencing of 18 Mb of wheat chromosome 3B revealed the presence of an unexpected high amount of nonsyntenic genes compared to the related grass genomes: rice, *Brachypodium* and sorghum. These genes have been inserted recently to a new locus on 3B, after the divergence of wheat and Brachypodium, 20-30 million years ago. This study aims at characterizing the origin and the number of copies of these genes in the hexaploid wheat genome. Thus, 53 nonsyntenic genes and 81 syntenic genes have been used to study the impact of gene duplications on wheat genome evolution. A bioinformatics approach has been developed in that purpose. Based on 454 reads available from a 5x whole genome shotgun performed on the wheat genome, the *BlastMap* pipeline has been developed in order to estimate the number of copies of each candidate gene. Results showed that nonsyntenic genes are duplicated at a higher degree than syntenic ones within the wheat genome. In total, 2/3 out of them are present in 6 copies or more, and 13% are duplicated in more than 18 copies, suggesting the role of transposable elements in gene capture and mobilization. This work revealed the amazing high level of interchromosomal duplications in the *Triticeae* lineage compared to closely related species.