Habilitation à Diriger des Recherches

François Sabot

Université Montpellier II

Jury Composé de :

- Dr Frédérique Pelsy, DR INRA, Rapportrice
- Dr Hadi Quesneville, DR INRA, Rapporteur
- Dr Eric Bonnivard, MdC Université Pierre et Marie Curie, Rapporteur
- Dr Pierre Capy, Pr Université Paris-Sud, Examinateur
- Dr Jacques David, Pr Montpellier Sup'Agro

La Génomique, socle nécessaire pour comprendre l'évolution La connaissance des génomes a fait en 20 ans un bon considérable, améliorant ainsi la compréhension des mécanismes de l'évolution des plantes, de la biodiversité, de l'amélioration variétale, et de la connaissance des éléments transposables. Nous savons maintenant que le génome végétal est un monde rempli d'océans d'éléments transposables, avec de petites îles géniques disséminées de-ci, de-là. Ces océans peuvent fortement influer sur les gènes, sur le phénotype global de l'organisme, voire même de l'environnement local (principe du phénotype étendu, avec par exemple les nudivirus chez les guêpes). De même, les reséquençages massifs dans de nombreuses espèces différentes, animales ou végétales, microbiennes, champignons, ou autres organismes inclassifiés/inclassifiables, ont permis de se rendre compte que les individus de la même espèce certes se ressemblent au niveau phénotypique, se croisent, etc... mais que leur contenu génomique diffère parfois fortement même entre individus proches. C'est l'ère de l'études des variants, de type SNP (*Single Nucleotide Polymorphism*), InDel, mais aussi et surtout structuraux : recombinaisons, larges insertions & délétions, transpositions, nouvelles séquences... Le génome d'une espèce ne se résume plus à une seule référence, et *Arabidopsis thaliana* n'est plus le modèle, mais plutôt l'exception, pour la génomique végétale. Dans le même temps, la découverte de l'épigénétique, des méthylations aux petits ARNs, a permis de découvrir tout un cycle de régulation supplémentaire, apportant son lot de variants et de contrôle, en particulier sur les éléments transposables.

Enfin, l'étude des éléments transposables a mis à jour les inter-relations complexes entre les éléments autonomes et les non-autonomes, qui les parasitent. Cette relation semble de plus s'étendre à des mécanismes évolutifs directement liés à l'évolution de l'organisme-hôte. C'est l'apparition de l'écologie génomique, où le génome-hôte est un écosystème, les différentes familles d'éléments les espèces le peuplant, avec chacune des insertions considérée comme un individu.

Et moi dans tout ça ? Cela fait maintenant 14 ans que je suis entré dans le monde de la Recherche végétale, depuis mon arrivée en DEA au laboratoire Amélioration et Santé des Plantes de l'INRA de Clermont-Ferrand. Au cours de mon DEA et de ma thèse, je me suis intéressé aux éléments transposables au sein des génomes des blés, en tant que marqueurs moléculaires d'abord puis comme composants principaux des génomes, et probables acteurs de la régulation des méïoses polyploïdes. Ensuite je suis allé travailler sur l'élément *BARE*-1 et son partenaire non-autonome *BARE*-2 au *Plant Genomics Laboratory* de l'Université de Helsinki, où j'ai conforté mes connaissances en biologie fonctionnelle des éléments transposables. Depuis mon recrutement à l'IRD, je suis plus sur les analyses en génomique massive des éléments transposables, avec un appui majeur de la bioinformatique. Au cours de cette période, j'ai pu étoffer mes compétences scientifiques et techniques, et évoluer dans différents laboratoires qui m'ont donné la possibilité de m'exprimer au mieux de mes compétences et de mes envies professionnelles, ce dont je les remercie.

Je suis allé de plus en plus loin sur les éléments transposables, mais aussi sur le génome au global. Je m'intéresse maintenant à l'écologie génomique, et à son impact sur le phénotype. Tout cela bien sur dans un contexte de soutien aux Pays et Partenaires du Sud, ce qui me donne un accès à énormément de matériel et de compétences de haut niveau sur des organismes très intéressants, comme les riz Africains, sauvages et cultivés.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Activités de Recherches Antérieures

1.1 Thèse de Doctorat - Clermont-Ferrand

J'ai effectué ma Thèse au sein du laboratoire ASP (Amélioraion et Santé des Plantes) de l'INRA de Clermont-Ferrand, dans l'équipe Génome, sous la direction de Michel Bernard, entre Septembre 2001 et Décembre 2004. J'avais auparavant effectué mon DEA dans ce même laboratoire sur l'utilisation des éléments transposables comme marqueurs moléculaires dans les blés.

1.1.1 Introduction

Le Blé - importance économique et culturelle Le blé tendre *Triticum aestivum* L., de la famille des *Poaceae* est la première céréale alimentaire en France, et une des trois premières cultures mondiales, avec le riz et le maïs. Cette espèce hexaploïde de génome ABD est issue d'une allopolyploïdisation suite au croisement de deux autres espèces, le blé dur tétraploïde *Triticum turgidum* donneur des génomes A et B, et l'herbe à chèvre diploïde *Aegilops tauschii* comme donneuse du génome D. Le blé dur est lui-même issu d'une allotétraploïdisation *via* le croisement de *Triticum urartu* donneur du génome A, et d'une espèce encore inconnue proche de *Aegilops speltoides*, donneuse du génome B (Figure 1.1) [Chantret et al., 2005].

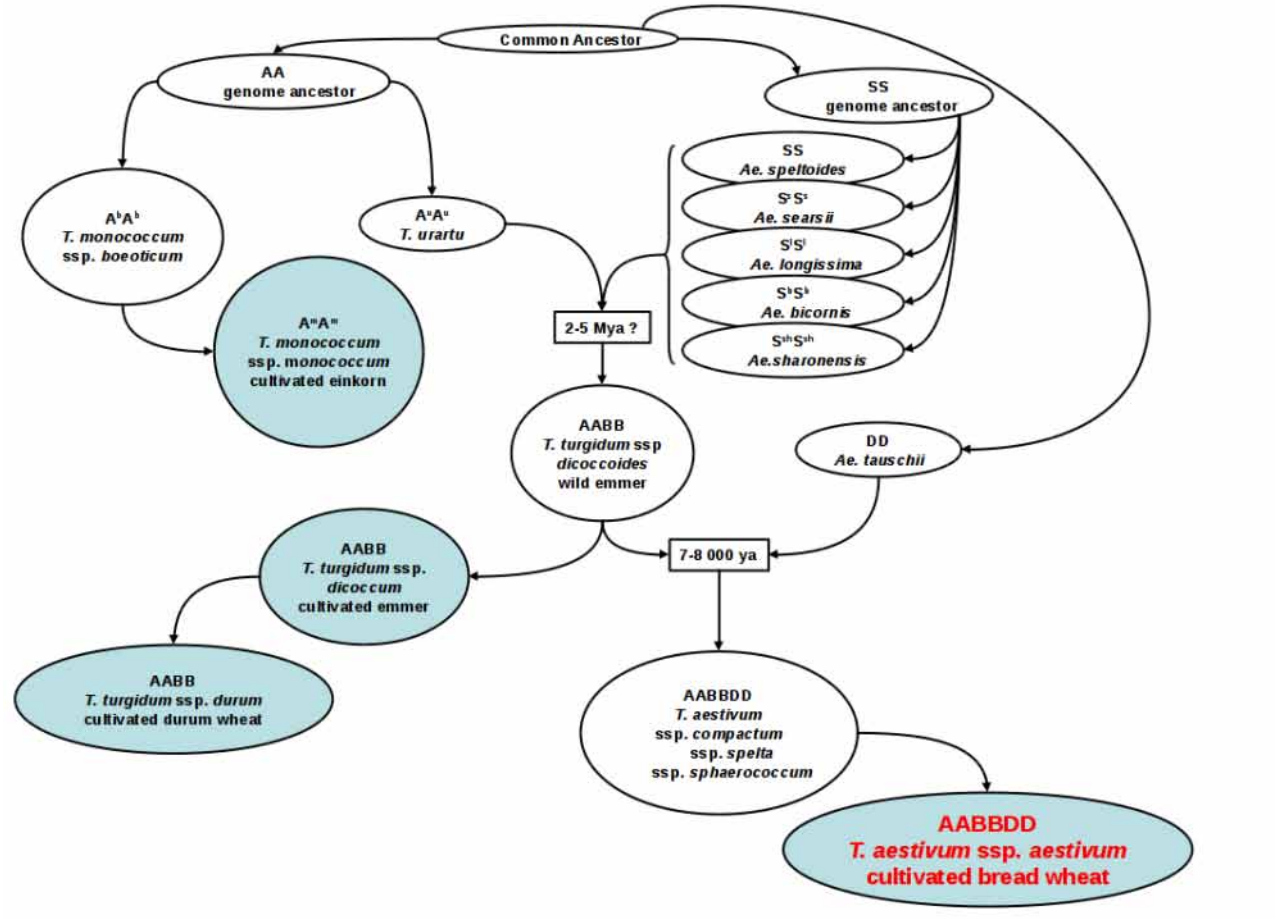

FIGURE 1.1 – Histoire évolutive du blé tendre. Sur fond blanc apparaissent les espèces sauvages, sur fond bleu les cultivées. *Mya= Millions years Ago.*

D'un point de vue économique, le blé tendre, qui sert à faire les farines pour les pains et les pâtisseries, est la première céréale française, la France étant un des 5 premiers producteurs mondiaux. Le blé dur, utilisé dans la fabrication des pâtes italiennes, et avec une production française beaucoup plus faible, est surtout cultivé en Europe du Sud et de l'Est. Enfin, le blé dit "petit épautre", de l'espèce *Triticum monococcum* voisine de *Triticum urartu*, est cultivé de manière confidentielle pour

entrer dans la fabrication de certaines farines multiples, en France principalement au pied du Mont Ventoux. Cette tribu de céréales, les *Triticeae* ont donc une très grande importance économique. De plus, c'est l'une des céréales du Croissant d'Or d'Asie Mineure, et son évolution est intimement liée à celles des cultures humaines. En effet, si l'apparition du blé dur il y a environ 3 millions d'années précède celle de l'agriculture, l'origine du blé tendre se confond avec la domestication des premières espèces agronomiques, il y a 10 à 12 000 ans. Enfin, il existe de nombreuses espèces polyploïdes dans cette tribu, et des versions diploïdes stricts comme polyploïdes existent pour quasiment tous les types génomiques répertoriés dans cette tribu (de A à V).

Sujet de Thèse Mon sujet de thèse se situait dans l'analyse évolutive globale de cette tribu des *Triticeae*, et plus précisément de l'évolution des différentes familles d'éléments transposables au cours des différentes spéciations et polyploïdisations qui ont eu lieu dans cette famille. Pour cela, j'ai utilisé deux approches, une dite *macro-génomique* (au niveau des génomes entiers) et une dite *micro-génomique* (sur une zone restreinte).

1.1.2 Approche Macro-génomique

Dans cette étude, j'ai utilisé une approche d'hybridation soustractive des génomes, *via* la RDA (*Representational Difference Analysis*, [Lisitsyn, 1995]), pour accéder aux séquences spécifiques d'un sous-génome (ou groupe de sous-génomes) par rapport à un autre.

La RDA : méthodes et limites Le principe est une soustraction *via* digestion et PCR de séquences communes entre le *Driver* (génome à éliminer) et le *Tester* (génome à conserver). Cela permet, après clonage, validation et séquençage, d'identifier les séquences spécifiques du *Tester* par rapport au *Driver*.

Dans le cadre de ma thèse, j'ai optimisé la méthode pour accéder aux éléments transposables spécifiques, en utilisant des sets d'enzymes insensibles à la méthylation, et en applicant une vérification post-sélection avec deux hybridations parallèles en *dot-blot*, l'une avec l'ADN total du *Driver* (contrôle négatif), l'autre avec l'ADN total du *Tester* (contrôle positif). Sans cette validation supplémentaire, 80% des séquences extraites resteraient des faux positifs non-spécifiques.

Ce travail a donné lieu à une publication en 2004 dans *Plant molecular Biology Reporter* [Sabot et al., 2004].

Résultats de la méthode Macro-Génomique Après avoir effectué plusieurs analyses de soustractions pour différents couples *Tester-Driver*, j'ai pu identifier entre 4 et 60 clones liés aux éléments transposables, représentant entre 3 et 20 sous-familles différentes (Table 1.1).

Couples	Clones RDA	Clones Positif	Clones ET	Autres Séquences	Redondance (ET)
ABD-AB	428	65 (15.2%)	32 (49.2%)	Inconnu	16 groupes (6)
ABD-D	96	5 (5.2%)	4 (NA)	Inconnu	4 groupes (3)
A-S	624	86 (16.4%)	49 (57%)	18S rRNA, *housekeeping*...	11 groupes (6)
ABxD-ABD	610	218 (35.7%)	60 (38.5%)	5S rRNA, répétitions...	38 groupes (20)

TABLE 1.1 – Résultats des analyses RDA sur différents couples de génomes. Le couple représente le *Tester* - le *Driver*. Le génome ABxD est celui du blé hexaploïde synthétique.

Certaines séquences ont ensuite été utilisées en *FISH* (*Fluorescent In Situ Hybridization*) pour valider et étudier leurs localisations chromosomiques (Figure 1.2).

Grâce à cette analyse, j'ai pu mettre en évidence une grande différence entre sous-familles d'éléments de type *Gypsy* entre les sous-génomes AB des blés durs et tendre, ainsi que des évènements de transpositions/envahissements lors des spéciations par polyploïdisation.

1.1.3 Approche Micro-génomique

Le Locus *Ha* et les données disponibles Le locus *Ha* régit la dureté des farines de blé, *via* l'action des puroindolines A et B. Ces deux gènes sont présents dans les génomes des espèces diploïdes

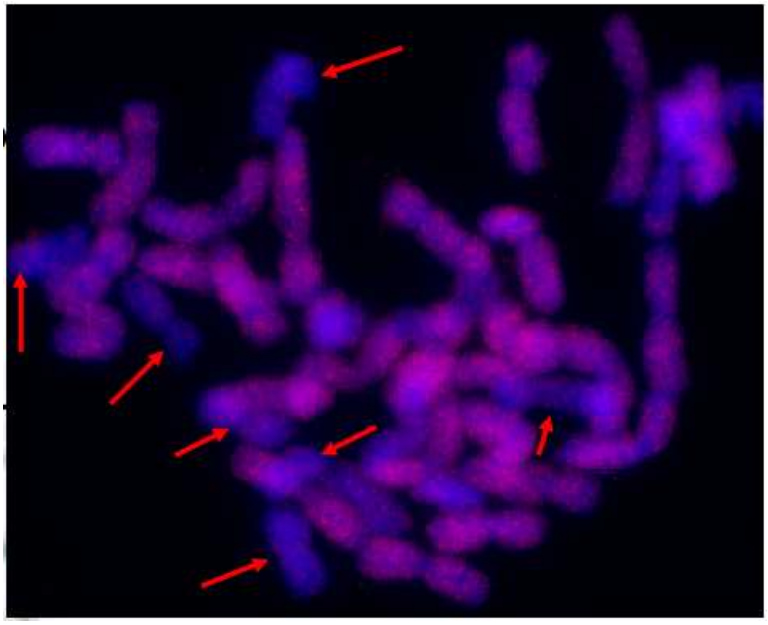

FIGURE 1.2 – Localisation chromosomique d'une sous-famille de *Fatima* (Rétrotransposon à LTR de type *Gypsy*) spécifique du génome A par rapport aux génomes S, sur le génome ABD du blé tendre. Les flèches rouges indiquent les chromosomes non marqués, du génome D.

de tout type (comme *T. urartu* ou *Ae. tauschii*), mais ont disparu dans le blé dur *T. durum*, qui a une farine dite dure. Il faut bien noter que cette élimination a eu lieu dans les deux sous-génomes, *a priori* de manière indépendante ! L'adjonction du génome D dans le blé tendre *T. aestivum* a permis de récupérer ces gènes, et donc une farine tendre. Jusqu'en 2003-2004, aucune information n'était disponible sur les mécanismes moléculaires liés à ces éliminations. Un premier séquençage a été réalisé sur un unique clone sélectionné d'une banque de BACs (*Bacterial Artificial Chromosome*) de *T. monococcum* (génome AA) à partir d'une sonde de la GSP-1 (*Grain Softness Protein-1*, gène localisé à très courte distance des purodindolines). Cette séquence a été analysée pour son contenu génique et non-génique ; j'ai effectué à cette occasion ma première annotation d'éléments transposables sur séquences de grande taille. Suite à l'analyse de ce locus *Ha* non délété, une analyse de plus grande envergure a été faite sur 8 BACs homéologues de différentes origines : les 3 sous-génomes du blé tendre (A, B et D), les deux du blé dur (A et B), le génome de *Ae. tauschii* (génome D), le génome de *Ae. speltoides* (génome S, ancêtre du B), et bien sur les données de *T. monococcum* déjà disponibles. Suite à l'analyse génique et non-génique de ces BACs, tous ancrés par la même séquence GSP, nous avons mis en évidence une excision liée à une recombinaison non-homologue des gènes de purodindolines A et B dans les deux sous-génomes A et B des blés polyploïdes (Figure 1.3), avec comme pieds 2 types d'éléments différents.

L'analyse proprement dite de ce locus a été publié en 2004 dans *MGG* pour les données de *T. monococcum* [Chantret et al., 2004] et en 2005 dans *Plant Cell* [Chantret et al., 2005] pour l'analyse sur les 8 BACs.

Nous avons ensuite continué le travail sur l'intégralité des séquences disponibles de ces BACs homéologues (Figure 1.4), de manière à analyser plus finement la microcolinéarité et l'évolution moléculaire de ces séquences, au niveau génique. Nous avons mis en évidence de fortes variations en termes de pseudogénisations dans une famille de gènes dupliqués en tandem *via* des duplications/délétions successives, ainsi que des variations importantes en termes de séquence, mais sans aucune liaison de proximité (deux gènes successifs n'avaient pas le même patron de divergence).

Ce travail a été publié en 2008 dans *Journal of Molecular Evolution* [Chantret et al., 2008].

Découverte de nouveaux éléments Lors de l'analyse de ce locus, j'ai identifié deux éléments ayant chacun une structure étrange, le *Morgane* et le *Veju_L*.

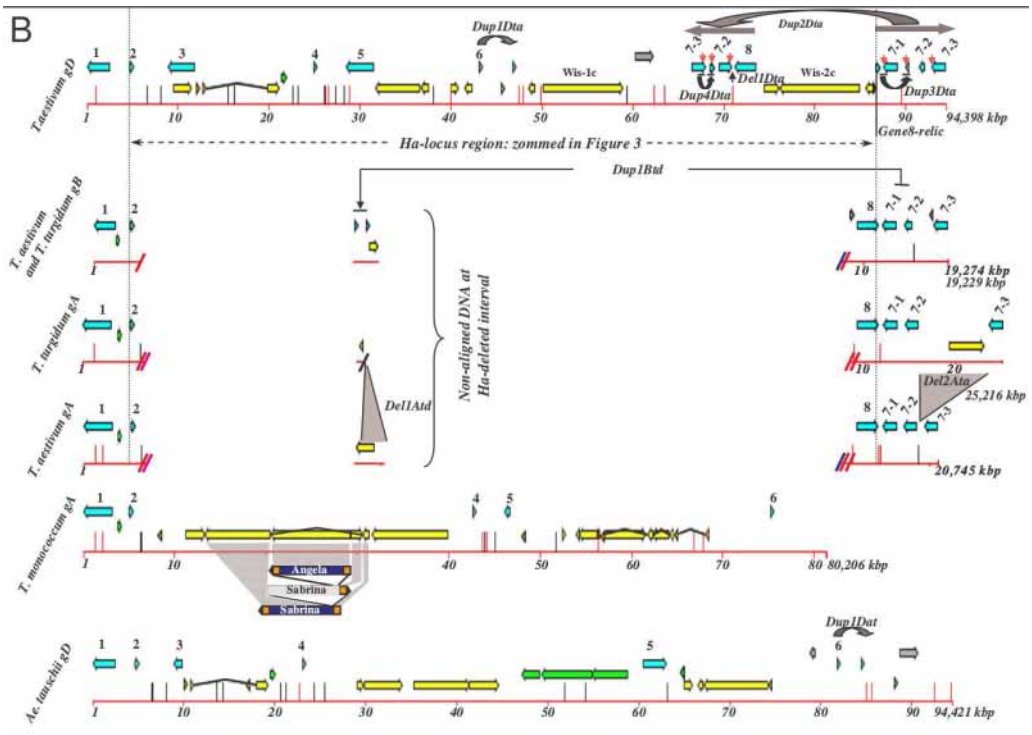

FIGURE 1.3 – Analyse de l'élimination du locus *Ha* dans les blés polyploïdes. De Chantret et al, 2005, Plant Cell

Veju_L L'élément *Veju* est un *TRIM* (*Terminal Repeat In Miniature*), petit rétrotransposon à LTR de 500 bases, présent dans le génome des *Triticeae*. Au cours de l'analyse du locus *Ha*, j'ai identifié un élément identique à ce *TRIM* mais long de plusieurs kilobases (Figure 1.5).

Après une analyse fine en bioinformatique et des expériences au laboratoire, j'ai pu mettre à jour deux sous-familles de *Veju*, une courte et une longue, la longue ayant été obtenue à partir de la forme courte probablement par *template switching* durant la rétrotransposition. Ce travail a fait l'objet d'une publication en 2005 dans *Genetica* [Sabot et al., 2005b].

Morgane *Morgane* est une petit élément (1.5 kb) de type rétrotransposon à LTR, qui ne comporte pas de séquence codante. Seule la fin de sa séquence interne comporte un fragment de reverse transcriptase (RT) de type *Gypsy* (Figure 1.6).

Cet élément a été identifié comme une différence insertionnelle entre le génome D du blé tendre et celui de *Ae. tauschii*, dans une séquence génique. Ses petits LTRs et sa petite taille, combiné à la présence de fragments de RT, en font un élément de choix pour comprendre les mécanismes conduisant d'un élément complet à un *TRIM*. Ce travail sur *Morgane* a donné lieu à un article plublié en 2006 dans *Genetica* [Sabot et al., 2006b].

Standardisation de la méthodologie d'annotation Suite à ces analyses d'éléments transposables, j'ai réannoté sur les mêmes critères standardisés les éléments transposables de la totalité des grandes séquences disponibles à l'époque sur les *Triticineae* (orges et blés), soit 5.1 Mb d'ADN répartis en 26 séquences. En utilisant différents outils et approches, j'ai pu augmenter de plus de 50% l'annotation en éléments transposables, et réaffiner les annotations existantes, pour ces différentes séquences.

Grâce à ces annotations standardisées, j'ai pu observé des différences entre les différents types de génomes en terme de composition en différents type d'éléments (Figure 1.7), ainsi qu'une insertion préférentielle des MITEs (*Minitature Inverted Repeats Elements*) à proximité des gènes.

Ce travail a été publié en 2005 dans *MGG* [Sabot et al., 2005a].

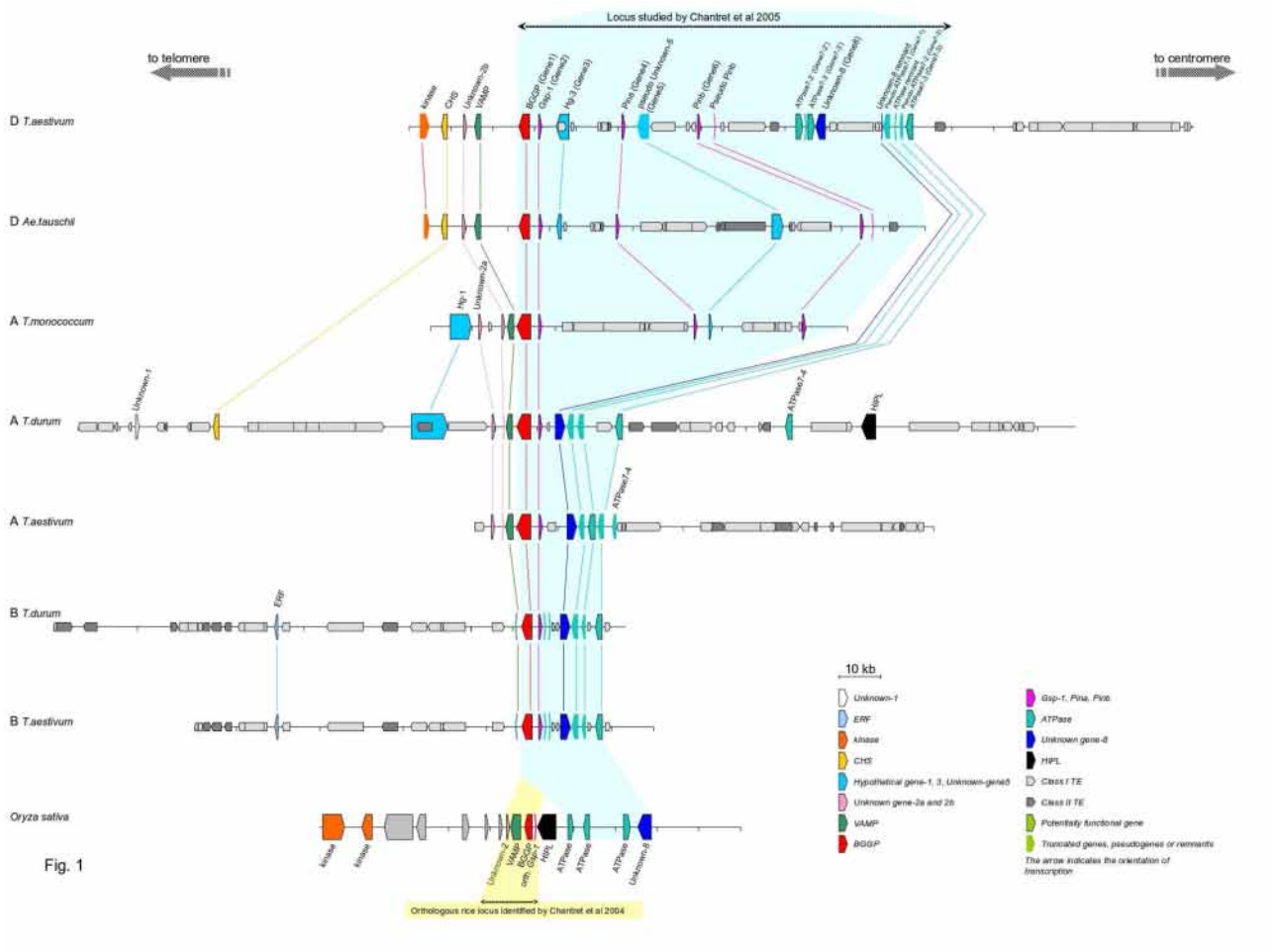

FIGURE 1.4 – Analyse de l'élimination du locus *Ha* et de la microcolinéarité les blés polyploïdes. De Chantret et al, 2008, *JME*

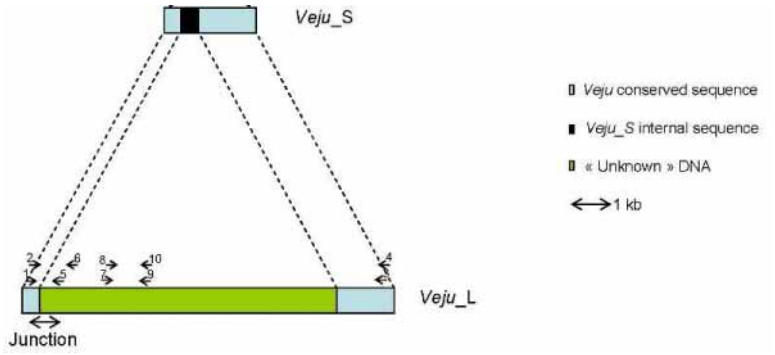

FIGURE 1.5 – Structure de l'élément *Veju*, forme courte (_S) et longue (_L).

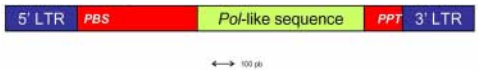

FIGURE 1.6 – Structure de l'élément *Morgane*.

Résultats de la méthode Micro-génomique L'analyse de ces différentes séquences (homéologues ou non) m'a permis de mettre en évidence des différences structurales (entre autres en terme de composition en éléments transposables) entre les différents types de sous-génomes de blé. Ainsi le

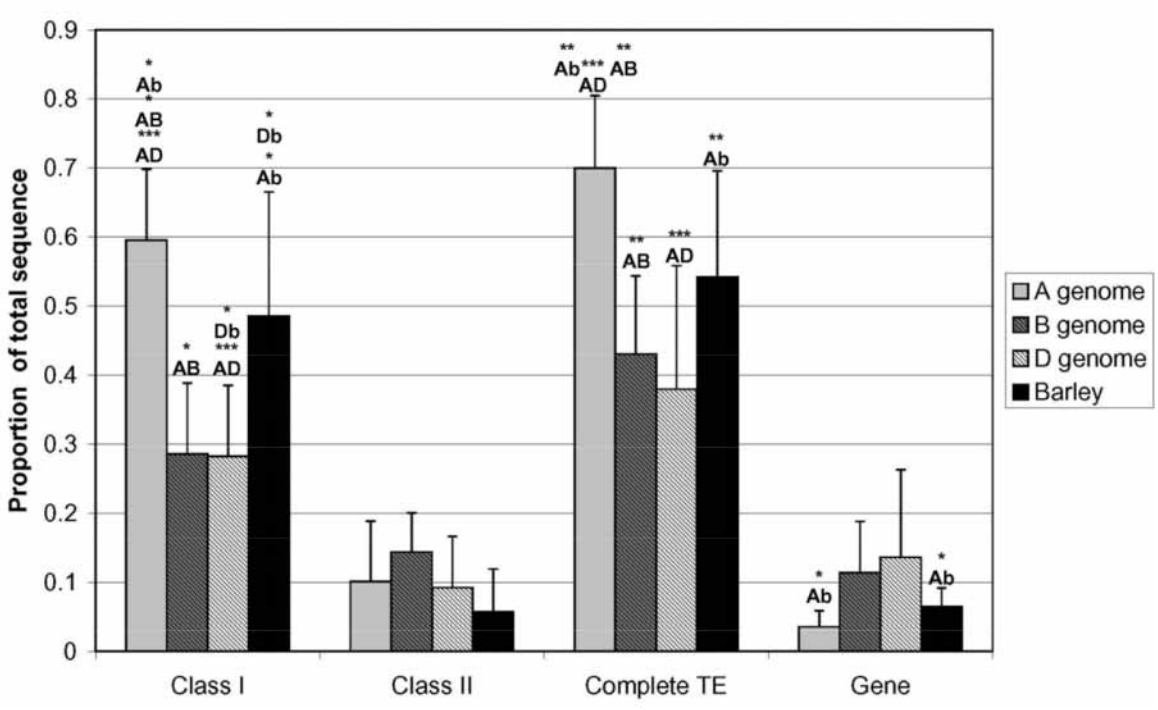

FIGURE 1.7 – Composition différentielle des différents types de sous-génomes en fonction des types d'éléments transposables.

génome A semble porter plus de rétrotransposons à LTR et autres éléments que les autres génomes, alors que le génome B être plus riche en gènes. De plus, l'analyse des gènes du locus *Ha* a aussi permis de voir que le concept de microcolinéarité est très souvent mis à mal dès que le niveau de définition et de précision augmente.

1.1.4 Conclusion des travaux doctoraux

Lors de ma thèse, j'ai pu mettre en évidence des familles et sous-familles spécifiques d'élements transposables dans les blés ou dans certaines espèces de blés. Ceci m'a permis de proposer à cette époque (Décembre 2004) une première ébauche des évènements transpositionnels durant l'évolution des blés (Figure 1.8).

A chaque étape de spéciation et de polyploïdisation semble correspondre une activité d'éléments transposables, soit par des amplifications de sous-familles spécifiques (comme les différentes sous-familles d'*Angela*), soit carrément par l'apparition de nouvelles familles (*Veju_L* et *Morgane* par exemple). Ces différents *bursts* pourraient être à l'origine de l'identité chromosomique spécifique des différents sous-génomes des blés polyploïdes, permettant des méïoses normales et le maintien de l'espèce, chez les blés. La différence temporelle de condensation chromosomique en métaphase permet une association en divalents correcte, *via* l'activité du locus *Ph* [Colas et al., 2008], mais aucune démonstration n'est actuellement satisfaisante dans la mise en oeuvre de cette condensation différentielle. Il est envisageable que les différences sous-génomiques en termes d'éléments transposables soient donc responsables de cette condensation, à la manière d'une serrure reconnue par une clef spécifique.

Ceci est bien sur une ébauche basée sur les données de l'époque. Il serait bon de réactualiser ces analyses avec l'aide des données obtenues grâce aux différents séquençages actuels sur les génomes des blés.

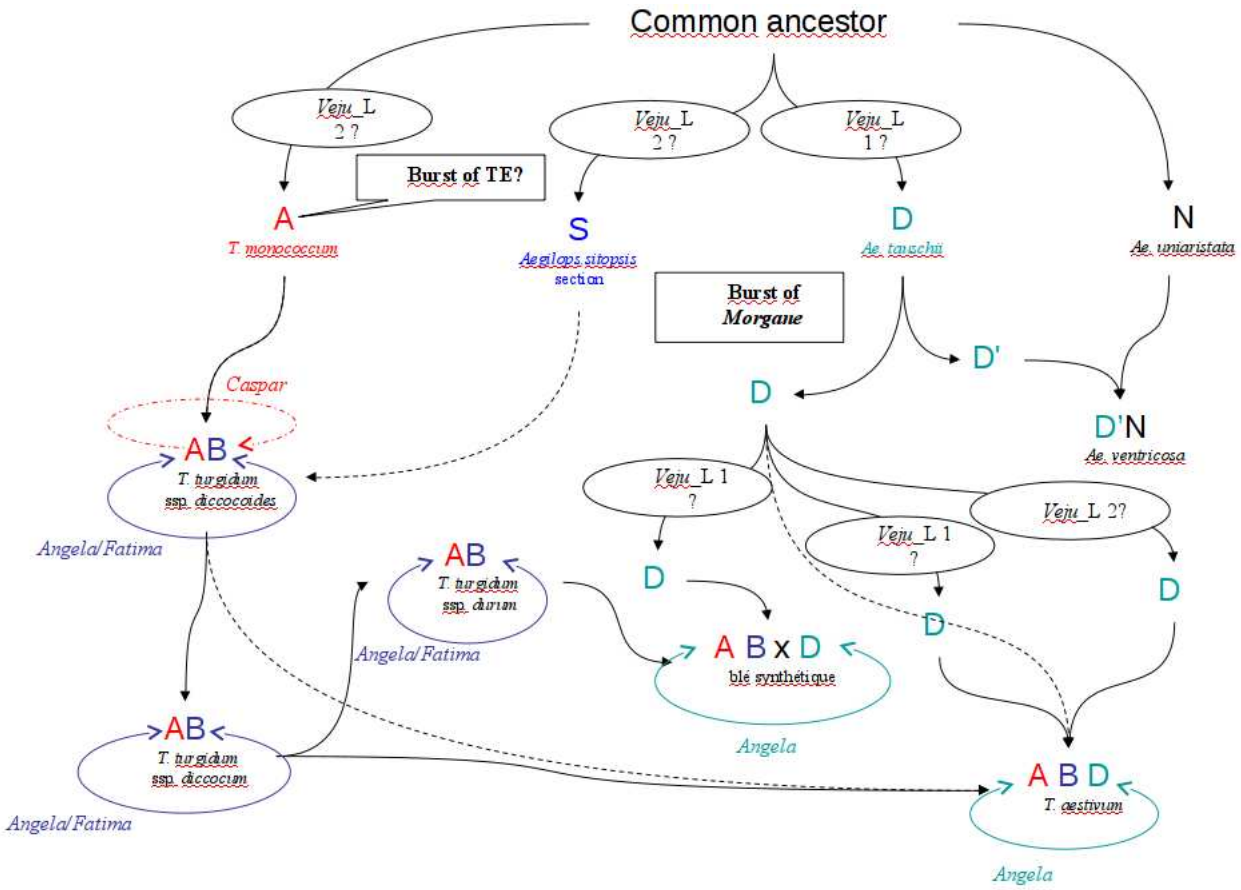

FIGURE 1.8 – Histoire évolutive supposée des blés en relation avec les activités des éléments transposables.

1.2 Activité Post-Doctorale - Helsinki

Une fois ma thèse soutenue, je suis allé au *Plant Genomic Laboratory* de l'Université d'Helsinki/Institut de Biotechnologie, dirigé par Alan Schulman, de Janvier 2005 à Octobre 2007.

1.2.1 Introduction

Le laboratoire travaillait essentiellement sur l'élément *BARE*-1 (Figure 1.9) de l'orge cultivé *Hordeum vulgare*. Ce rétrotransposon à LTR de type *Copia* a la particularité d'être présent en un très grand nombre de copies dans son génome-hôte (120 000 copies estimées, donc 40 000 entières, [Kalendar et al., 2000]).

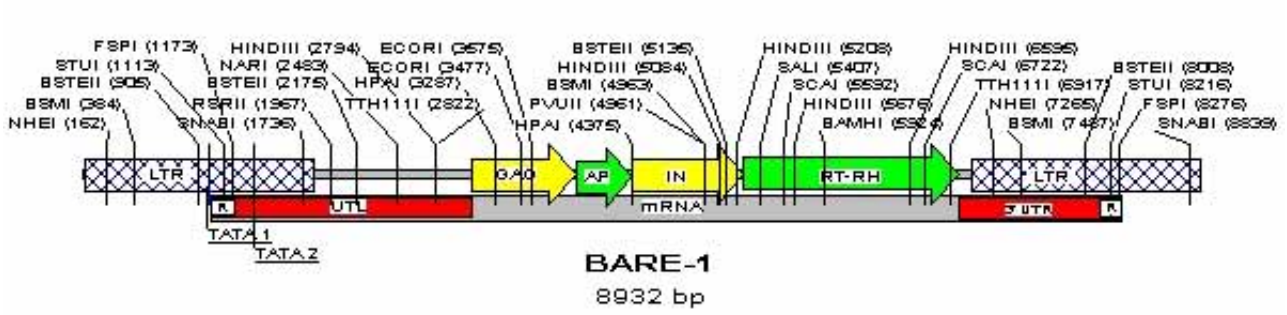

FIGURE 1.9 – Structure de l'élément *BARE*-1 de l'orge, avec les LTR flanquants, ainsi que les deux ORFs *Gag* et *Pol*. Sont figurés aussi les sites de coupure enzymatique. De Schulman.

Cet élément m'a servi de modèle pour commencer à développer les analyses sur les relations entre éléments autonomes et non-autonomes. En effet, cette famille d'éléments est formé de deux sous-familles, *BARE*-1 et *BARE*-2

1.2.2 *BARE*-1/*BARE*-2, une histoire de GAG

L'élément *BARE*-1 se multiplie *via* un mécanisme de rétrotransposition classique pour un rétrotransposon à LTR (Figure 1.10). Au sein du laboratoire, nous travaillions plus précisément sur la partie de l'expression et de l'encapsidation des ARNs de rétrotransposon, en utilisant la GAG de *BARE*-1.

Activité de la GAG de *BARE-1* sur son ARN en propre Les travaux sur le *VIH* et d'autres lentivirus proches comme le *SIV* ont mis en évidence une spécificité très forte de l'ARN matrice de l'élément pour sa propre GAG, *via* l'action du *PSI* (*Packaging SIgnal*). Cette reconnaissance tridimensionnelle assure un *packaging* spécifique à plus de 98% de ce type de virus [Dannull et al., 1994]. Partant de cette idée, avec Marko Jääskelaïnen, un des doctorants du laboratoire, nous avons mis au point une approche triple-hybride pour tester cette spécificité pour l'élément *BARE*-1 (Figure 1.11). Pour ce faire, j'ai cloné les 400 premières bases de l'ARN de l'élément, pour obtenir le *PSI* et le *DIS* (*DImerization Signal*) dans la même séquence, intégré cette séquence dans un vecteur, la séquence de la GAG dans un 2e, et enfin finalisé le 3e vecteur necéssaire à l'analyse. Et nous avons mis le tout dans une levure, fait des contrôles négatifs, attendu 3 jours, etc... *In fine*, comme le montre la Figure 1.11, toutes les interactions sont positives... Nous en avons conclu que la GAG avait une constante d'affinité très basse, et que sa sur-activation entraînait une activité non-spécifique.

Nous avons donc choisi une autre approche, *via* gradient de densité de saccharose. Ici, les composants d'un échantillon donné sont séparés à partir de leur masse et de leur densité : plus un objet est lourd ou complexe, plus il migrera vers le fond de la colonne. Après avoir récupéré des fractions différentielles (par mL), puis déposé sur gel non-dénaturant et effectué un *Western-blot* avec un anticorps anti-GAG, nous avons quantifié la GAG dans chaque fraction, et dans différentes conditions (Figure 1.12).

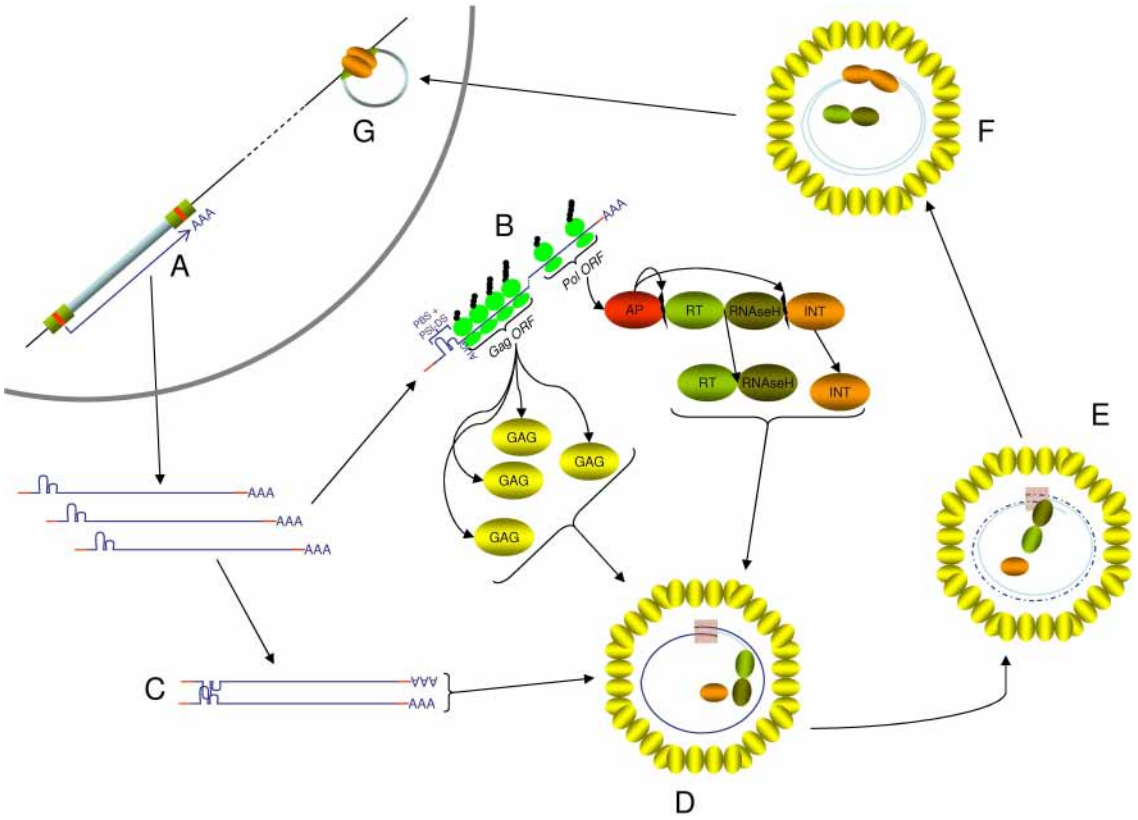

FIGURE 1.10 – Cycle de rétrotransposition des rétrotransposons à LTR. **A**) Transcription de type Pol II de l'ARN, du début de la région *R* du LTR 5' à la fin de cette même région sur le LTR 3', suivi d'un export dans le cytosol. **B**) Une partie du pool ARNm est traduit (après épissage ou non) en GAG et POL. Les GAG se polymérisent pour former une *VLP* (*Virus-Like Particle*) requise pour l'encapsidation des ARN matrices. La POL est ensuite séparée par une activité d'endoclivage de l'AP (*Aspartic Protease*) en RT-RH (*Reverse Transcriptase RNAseH*) et INT (*INTégrase*). **C**) les ARN matrices se dimérisent *via* l'action du *DSI*. **D**) les ARN dimérisés sont spécifiquement reconnus par leur GAG (action du *PSI*) et encapsidés dans la *VLP*. **E**) Le premier brin matrice ARN est transformé en cDNA simple-brin, puis est dégradé. **F**) le deuxième brin du cDNA est synthétisé. **G**) le cDNA double-brin est réinséré dans le génome-hôte après transfert dans le noyau grâce à l'association spécifique avec l'INT.

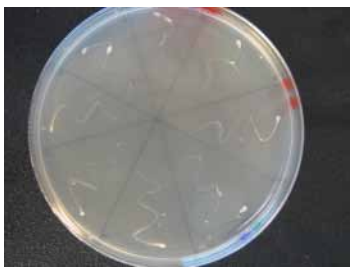

FIGURE 1.11 – Résultat de validation en triple-hybride de la spécificité de la GAG de *BARE*-1 sur son propre ARN.

Si l'on adjoint à la GAG un extrait d'ARN totaux de cal (où l'élément *BARE*-1 s'exprime fortement), on voit que la sédimentation a lieu plus loin (points rouges *vs* contrôle en noir), indiquant la formation d'un polymère en présence des ARNs (formation de la *VLP*). De même, en ajoutant moins de $\frac{1}{40}$ de cette quantité d'ARN sous forme d'ARN de *PSI* de *BARE*-1 uniquement (ligne verte), on note un accroîssement très important des formes polymériques. Cette formation est détruite en présence

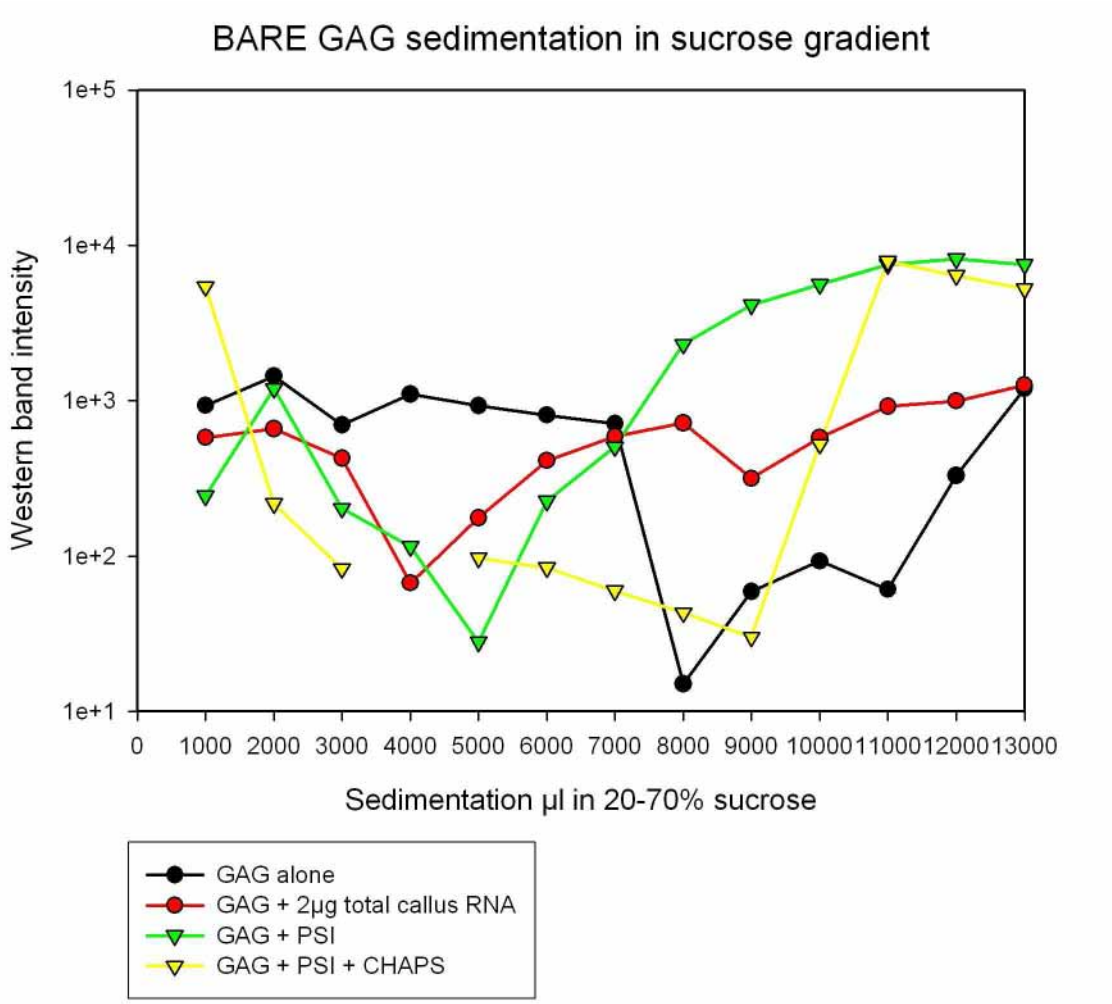

FIGURE 1.12 – Analyse de l'association entre la GAG et son ARN sur colonne de sédimentation. La quantité de motif *PSI* déposée correspond à moins de 50 ng d'ARN. De Jääskelaïnen.

de détergent (*CHAPS*), indiquant une association non-covalente. La GAG de *BARE*-1 s'associe donc de manière préférentielle et optimale avec son propre ARN, *a priori* a partir de la région situé entre la fin du LTR 5' et 400 bases en aval.

Parasitisme de *BARE-2* sur *BARE-1* pour l'encapsidation les *VLP* La sous-famille *BARE*-2 est issue d'une recombinaison entre les éléments *BARE*-1 et *Wis*-2, très proches (Figure 1.13). Cette recombinaison a produit une délétion spécifique de 8 nucléotides juste sur le site de début de traduction de l'ORF (*Open Reading Frame*) de la GAG (perte de l'ATG). Le second ATG se retrouve bien plus loin dans cet ORF, ne permettant pas de produire une GAG fonctionnelle. Cette délétion spécifique m'a permis de dériver des marqueurs PCR spécifiques de chacunes des deux sous-familles.

Bien que la zone interne de *BARE*-2 soit plus proche de *Wis*-2 que de *BARE*-1, les deux éléments *BARE* partagent les sites de régulation de la rétrotransposition (Figure 1.10) : le *PBS* (*Primer Binding Site*, nécessaire pour la formation du 1er brin de l'ADNc), l'extrémité 3' des LTRs (pour une association spécifique avec l'intégrase), ainsi que la zone du *PSI* (nécessaire pour la reconnaissance spécifique entre l'ARN du rétrotransposon à LTR et sa GAG) et du *DIS* (permettant l'association spécifique des deux matrices d'ARNs). Ces deux derniers signaux sont localisés juste après le LTR 5'.

L'analyse de la séquence des deux types de *BARE*, ainsi que de leur répartition génomique dans les différentes espèces d'orge et la quantité d'ARN exprimé pour chacun d'eux m'a permis de mettre en évidence un *burst* de transposition des *BARE*-2 dans l'orge cultivé et son ancêtre sauvage direct, en relation avec une surreprésentation de l'élément non-autonome en termes de copies et d'expression : le *BARE*-2, non-autonome, est deux fois plus présent que *BARE*-1, l'autonome.

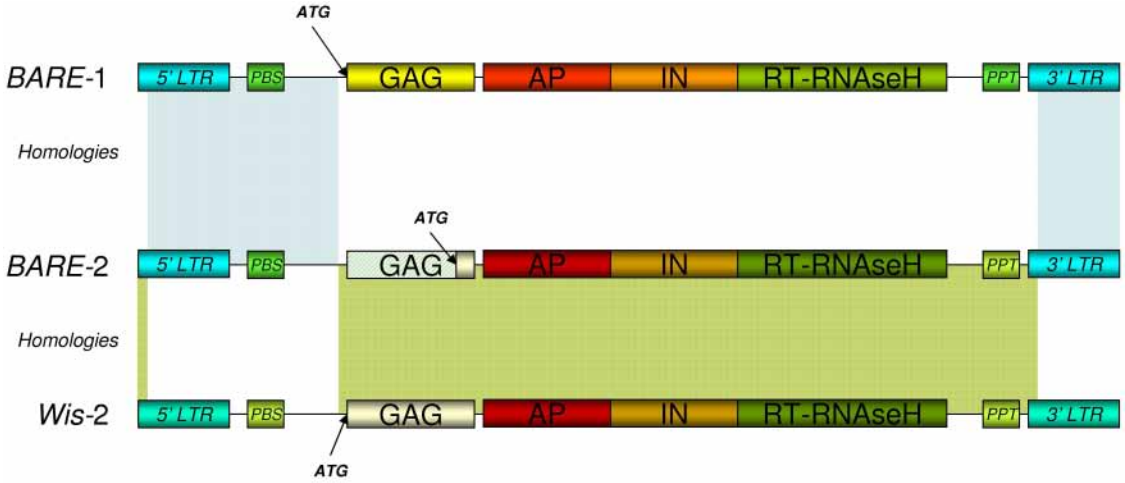

FIGURE 1.13 – Relations structurales entre les éléments *BARE*-1, *BARE*-2 et *Wis*-2.

Ce travail a été publié en 2007 dans *Gene* [Tanskanen et al., 2007].

En faisant la même expérience de sédimentation que précédemment avec la zone *PSI* de *BARE*-2, nous obtenons le même résultat d'association. Une expérience supplémentaire a été faite par Wei Chang, une autre doctorante du laboratoire. Après une extraction des protéine cytosolique, et un traitement à la DNase, l'echantillon débarrassé de toute forme d'ADN libre a été passé au CHAPS, dénaturant ainsi les protéines. Enfin, une simple PCR avec des adaptateurs spécifiques de *BARE*-1 et d'autres de *BARE*-2 (situés sur la zone de perte de l'ATG de la GAG, les mêmes que ceux que j'avais utilisé plus tôt) a été faite. Ca nous a permis de constater que notre échantillon contenait encore de l'ADN, mais uniquement celui de *BARE*-2 ! La PCR de *BARE*-1 n'a pas permis de récupérer des échantillons dans l'extrait. Ceci semble montrer que les ADNc de *BARE*-2 sont bien présents dans la *VLP* (protection à la DNase), et outrepassent ceux de *BARE*-1 dans l'encapsidation. Dans ce cadre, le mécanisme de parasitisme semble se rapprocher des inhibitions compétitives de type enzymatique, avec une plus grande disponibilité des ARNs de *BARE*-2 par rapport à ceux de *BARE*-1, qui eux sont en cours de traduction, complexés aux ribosomes.

1.2.3 Les autres éléments non-autonomes

***Sukkula*, celle qui a mordu la main...** D'autres éléments non-autonomes existent dans les gros génomes des céréales, comme *Sukkula*. Ce rétrotransposon à LTR de 12 kb ne comporte aucun ORF susceptible de coder pour sa machinerie de transposition [Kalendar et al., 2004]. Par contre, des analyses antérieures avaient montré une forte conservation de la zone immédiatement après le LTR 5', conservation associée à une forte information. Il semble donc que cette zone soit importante pour l'élément : elle comporte le *PSI* et le *DIS*. J'ai recherché *via* des outils d'analyse et de recherche de motifs, ainsi que par comparaison (manuelle) de nombreuses structures secondaires des ARNs les candidats possibles pour le partenaire autonome de *Sukkula*. A la suite de ces analyses, il s'avère que le partenaire le plus probable est *Erika*, un rétrotransposon à LTR de type *Gypsy* présent dans les *Triticeae*. Malheureusement, les copies connues de *Erika* dans l'orge étaient toutes fortement mutées ou délétées sur leurs séquences codantes. Le génome n'étant pas disponibles à ce moment-là, j'ai travaillé à mettre au point deux types d'outils moléculaires pour identifier les partenaires actifs

Clonage direct de séquence active La perte d'ORF correct des éléments *Erika* connus m'a conduit à générer un type de plasmide de clonage (Figure 1.14) spécialement étudié pour pouvoir cloner un produit PCR classique et permettre l'identification immédiate des clones portant un ORF fonctionnel (sans codon stop ou non-sens).

Ce plasmide, dérive du *pET 14b* existe en 3 versions, pour permettre un décalage de cadre de lecture. En pratique, le clonage se fait en amont de la kanamycine, avec un vecteur rapporteur de

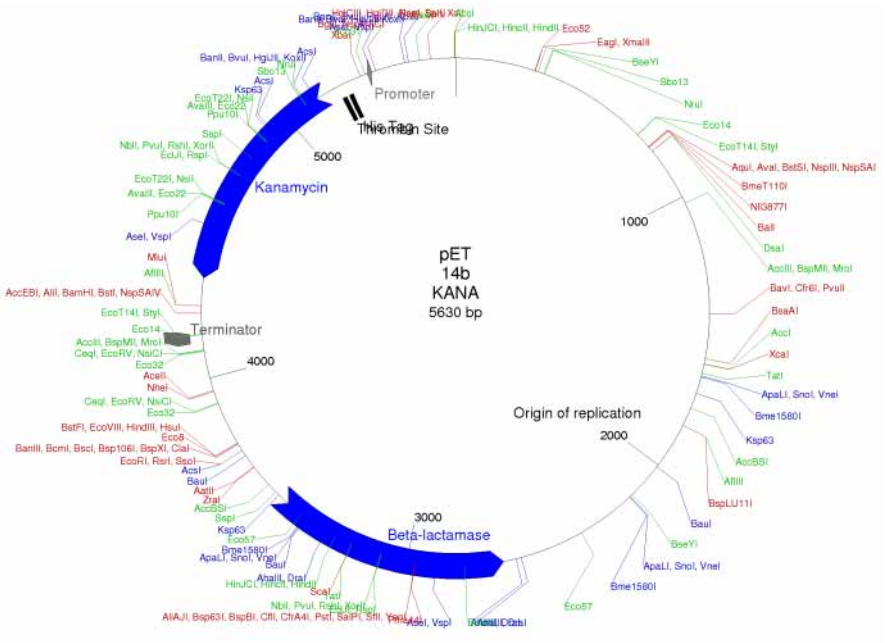

FIGURE 1.14 – Plasmide de clonage direct des produits PCR avec identification des séquences ORF fonctionnelles.

transformation (la β-lactamase). Si les bactéries transformées survivent en milieu kanamycine, la séquence clonée comporte un ORF putativement fonctionnel. Il est donc possible de trier parmi l'intégralité des séquences amplifiées en PCR par le même set de primers les ORFs actifs des autres, et ainsi de "pêcher" au sein d'une mare de séquences mutées la ou les quelques copies actives. Malheureusement, même si le plasmide était finalisé (pour le cadre +1), il ne m'a pas été possible de l'utiliser, mon post-doc prenant fin.

Détection d'interaction de structure secondaire ARN-ARN *in vivo* Dans la même idée d'identifier des partenaires actifs, il me fallait disposer d'un outil permettant de valider une association ARN-ARN en structure secondaire, *in vivo*. J'ai donc commencer à mettre au point un tel outil, basé sur la *GFP* et l'épissage en *trans* possible chez les eucaryotes (Figure 1.15).

Le plasmide *pA* a été fini, le *pB* pas complètement, et comme dans le cas du clonage d'ORF fonctionnel, mon post-doc touchant à sa fin, je n'ai pas pu les terminer ni les tester en conditions réelles.

Sukkula & Erika Le travail reste donc à finir, avec la validation du couple autonome/non-autonome. Si je devais reprendre ces aspects aujourd'hui, ma priorité porterait sur la finition des plasmides, de manière à valider expérimentalement nos données informatiques.

1.2.4 Les insertions complexes

Durant mon post-doc, je me suis aussi beaucoup intéressé à l'annotation de séquences d'éléments transposables. Dans ce cadre là, j'ai identifié des structures étranges d'insertion de rétrotransposon à LTR (Figure 1.16) : ces éléments étaient formé par une succession LTR-séquence interne-LTR-séquence interne-LTR.

Ce genre de structure était connu déjà chez *Arabidopsis* [Devos et al., 2002], mais avait été montré comme dérivant de recombinaisons post-insertions. Dans notre cas, les LTRs extérieurs étaient bordés des mêmes 5 bases (*Target Site Duplication - TSD*) et étaient beaucoup plus similaires l'un à l'autre que par rapport au LTR central (Figure 1.17). Ces deux signaux (*TSD* et identités des LTRs extérieurs) nous a conduit à imaginer un mécanisme de reverse transcription anormal, basé sur une encapsidation de non pas 2 molécules mais 4 molécules d'ARN matrices (Figure 1.18).

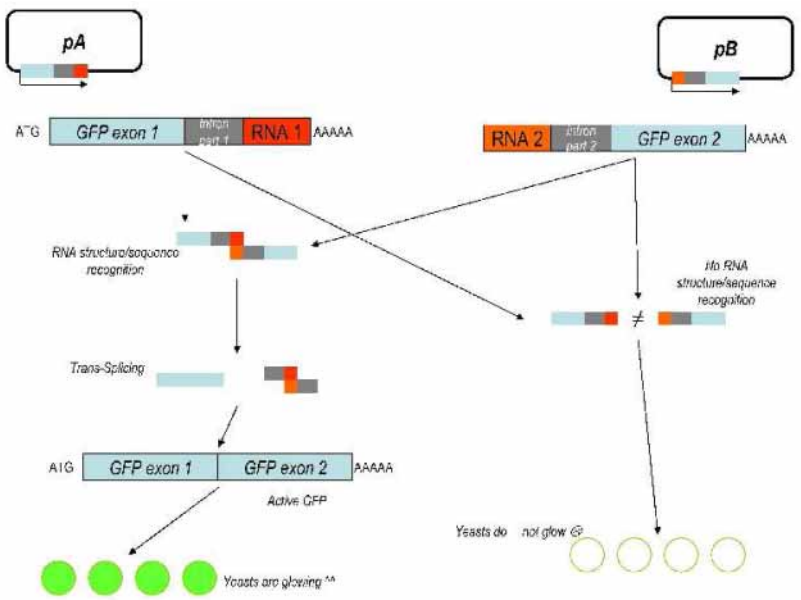

FIGURE 1.15 – Double système plasmidique de levure permettant une identification des associations ARN-ARN en structure secondaire *in vivo*. Si les deux ARNs se reconnaissent et interagissent, l'ARNm secondaire de la *GFP* peut être formé, et la levure émet une fluorescence verte.

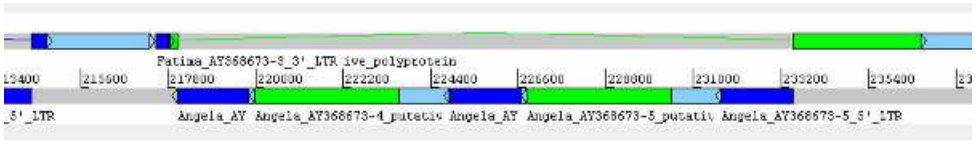

FIGURE 1.16 – Structure de type Complexe d'un élément *Angela*. On distingue bien les 3 LTRs en bleu foncé et les deux régions internes en bleu clair. En vert sont noté les séquences codantes.

Dans un tel système, la "boucle" du brin matrice se fait non pas entre les deux régions R du même élément, mais entre des régions R de deux éléments différents. Le saut de matrice suite au *strong stop* se fait alors sur l'autre ARN matrice, et non pas sur le même, permettant ainsi de former cette structure complexe (Figure 1.18).

Après avoir analysé plus de 1400 insertions anormales sur l'orge, le blé, le maïs et le riz, j'ai identifié un unique évènement de rétrotransposition complexe chez le maïs et le riz, et 2 chez les *Triticeae*. Un rapide calcul basé sur la taille des génomes par rapport à la taille des données analysées laisse supposer que près de 6000 évènements de ce type peuvent exister dans le génome des *Triticeae*. Il est intéressant aussi de noter qu'en théorie, de tels éléments peuvent être transcrits de manière complètement conventionnelle, *i.e.* du début de la région R d'un LTR en 5' jusqu'à la fin de la région R du LTR directement en aval.

Ce travail a été publié dans *BMC Genomics* en 2007 [Sabot and Schulman, 2007].

1.2.5 Nomenclature & Classification

Au cours de mon post-doc j'ai aussi été amené à travailler sur la nomenclature et la classification des éléments transposables eucaryotes. La problématique était simple : beaucoup de génomes complexes allaient être disponibles dans les 4 à 6 ans, et il n'existait pas de règles bien définies sur la manière d'identifier et de nommer des éléments transposables, particulièrement chez les plantes. J'avais fais partie d'un premier groupe de réflexion en amont du premier congrès de *RepBase* (Asilomar 2006) sur les TEs eucaryotes, congrès auquel je n'ai pu participer en personne. Après Asilomar 2006, un groupe de travail avait été défini sur ces aspects, normalement géré par Jerzy Jurka, mais l'intention est restée lettre morte... En Janvier 2007, au congrès *Plant & Animal Genomes* de San Diego, les consortiums

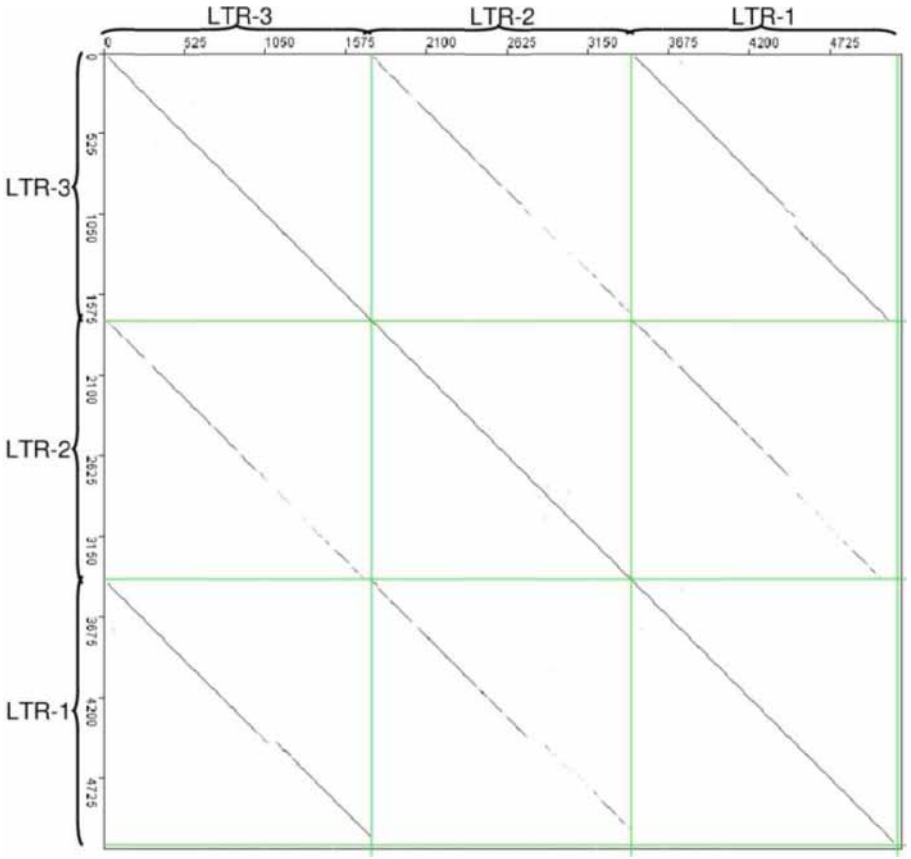

FIGURE 1.17 – *Dot-plot* des 3 LTRs du même complexe. On distingue bien la plus forte identité de séquence entre les deux LTRs extérieurs par rapport au central.

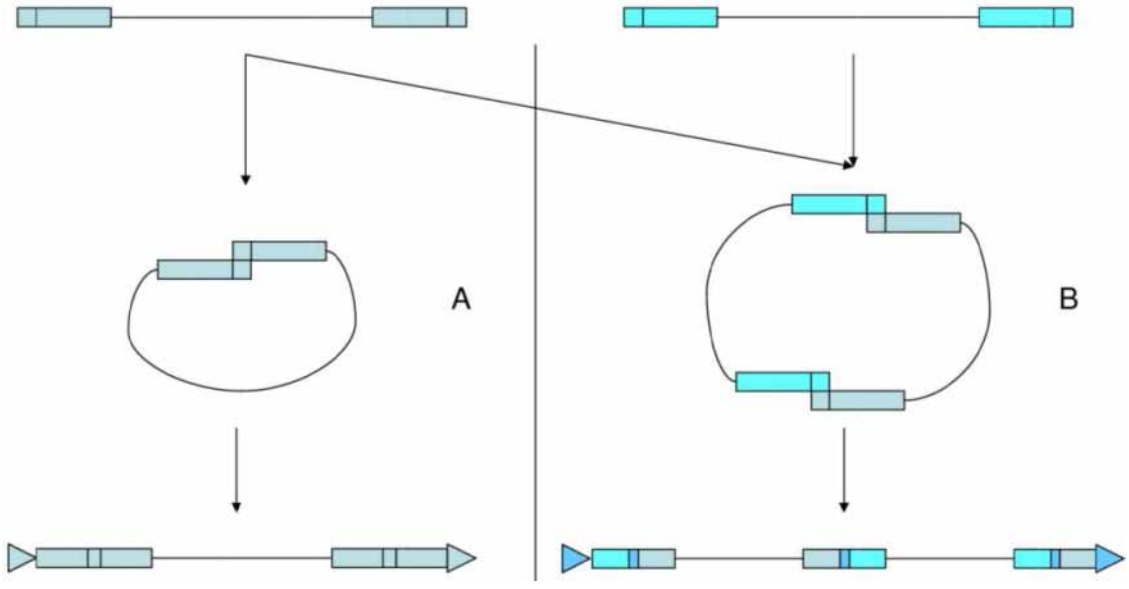

FIGURE 1.18 – Mécanisme de formation des insertions complexes. A) association et reverse transcription normale. B) association anormale et formation d'un complexe.

de séquençage du blé et de l'orge nous ont mis au pied du mur, réclamant des méthodes simples pour la classification, sous peine de mettre "*Repeat Regions*" sur toute séquence répétée se présentant. Avec Thomas Wicker et Alan Schulman, nous avons organisé une réunion sur place, avec plusieurs spécialistes plante, dans l'idée d'aboutir à un consensus. Suite à cette réunion, Thomas et moi avons

ressorti de nos cartons nos idées déjà proposées à Asilomar 2006, et les avons mises à jour. Le groupe initial formé à PAG a travaillé sur cette base, et sur conseil d'un éditeur de *Nature Reviews Genetics*, nous avons contacté des spécialistes animaux/champignons (Aurélie Hua-Van et Pierre Capy) pour se joindre à nous dans ce travail.

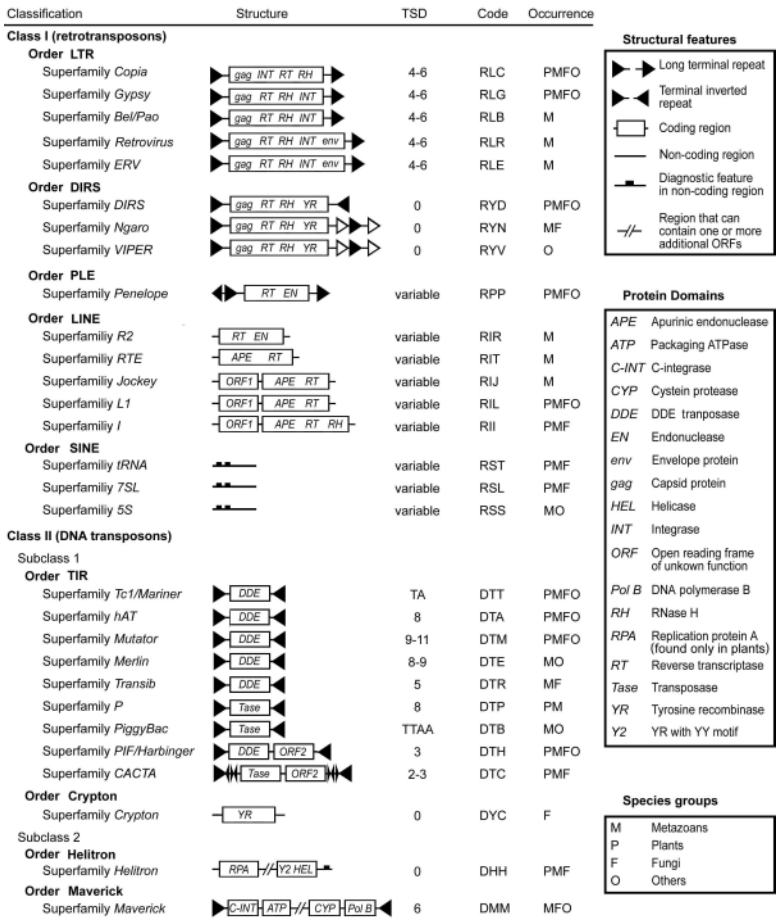

FIGURE 1.19 – Proposition de classification pour les éléments transposables eucaryotes.

Il en est sorti, après de nombreux allers-retours et discussions, une proposition de nomenclature (Figure 1.19) basée à la fois sur les mécanismes moléculaires de transposition (intermédiaire ARN ou ADN, etc...) mais aussi sur la structure des éléments pour obtenir un niveau de définition sur la famille. Nous avons aussi mis au point une méthode facilement codable de classification dans les familles connues, le fameux "80-80-80" : si une séquence comporte 80% d'identité de séquence nucléique sur au moins 80% de sa longueur, avec une taille minimale de 80 bases, avec un élément déjà connu, alors cette séquence fait partie de la même famille. Ce score a été calculer pour obtenir le même type de résultat qu'en utilisant une approche *wet lab* en *Southern blot*, et donne d'excellents résultats sur les *Poaceae*, par exemple.

Ce travail a fait l'objet d'une publication en 2007 dans *Nature Reviews Genetics* [Wicker et al., 2007], ainsi que de deux réponses, dans ce même journal, en 2008 [Wicker et al., 2008b] et 2009 [Wicker et al., 2009].

1.2.6 Conclusion sur les travaux post-doctoraux

Au cours de mon post-doctorat, j'ai énormément appris sur les éléments transposables, et en particulier sur les rétrotransposons à LTR. Je me suis intéressé à leur mécanisme de rétrotransposition (qui n'a encore jamais été validé en entier à ce jour chez les végétaux), à leur évolution, et surtout aux interactions moléculaires et évolutives entre les éléments autonomes et les non-autonomes. Ces éléments semblent être des super-parasites très efficaces, en général même plus que leur partenaire autonome, en terme d'amplification au sein du génôme-hôte, d'encapsidation et de reverse transcription. Certains

ont même été identifiés avant leur partenaire autonome, et ont servi de modèle d'étude (exemple de *Tos17*). J'ai publié sur ce sujet deux revue, parue dans *Heredity* [Sabot and Schulman, 2006] et *Israel Journal of Ecology and Evolution* [Sabot et al., 2006a] en 2006, ainsi qu'un chapitre de livre en 2009 [Sabot and Schulman, 2009].

1.3 IRD - Équipe TEPG - Perpignan

1.3.1 Introduction

A la fin de ma période post-doctorale, j'ai été recruté fin 2007 comme Chargé de Recherches de 2e Classe à l'IRD, sur l'évolution du génome du riz. Suite à ce recrutement, j'ai été affecté au Laboratoire Génome et Développement des Plantes (LDGP) de l'Université de Perpignan, dans l'équipe "*Transposable Elements in Plant Genomes*" (TEPG), sous la direction du Pr Olivier Panaud. Je suis arrivé au laboratoire en Novembre 2007, pour rejoindre Montpellier en Juillet 2010. Au cours de cette période je me suis intégré dans des projets existants, en collaboration locale ou nationale (dont l'annotation des éléments transposables de Classe II dans le génome du Cacaoyer), et ai développé mes propres projets de recherches, tout en finalisant mes publications de post-doctorat.

1.3.2 Annotations fonctionnelles des éléments transposables

A mon arrivée au LGDP, j'ai mis en pratique de suite mes compétences en termes d'annotation fonctionnelle des éléments transposables sur un projet en cours (*Lullaby*), puis sur un projet d'une doctorante (Anne Roulin), et enfin sur le plus connu des éléments du riz Asiatique *Oryza sativa*, le rétrotransposon à LTR *Tos17*

Lullaby *Lullaby* est un rétrotransposon à LTR de la superfamille *Copia*, identifié par Nathalie Picault en 2008 comme étant très actif transcriptionnellement pendant l'embryogénèse et la régénération du riz, mais avec peu de néoinsertions *in fine*. *Via* une analyse classique utilisant *Pfam*,*ProSite* et*BLAST* sur des bases spécialisées, j'ai reconstruit l'organisation protéique de cet élément (Figure 1.20).

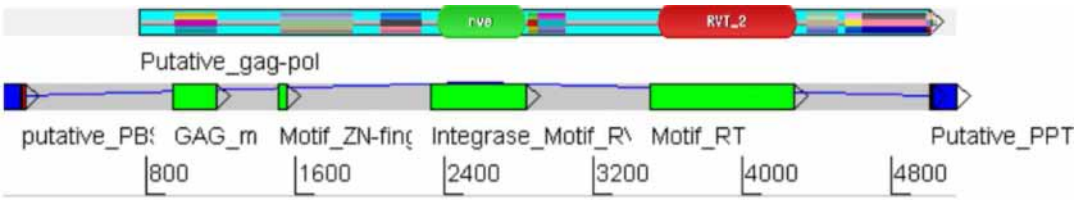

FIGURE 1.20 – Structure de l'élément *Lullaby*.

Curieusement, *Lullaby* est très similaire à *Tos17* sur une grande partie de sa séquence, mais pas sur la première région protéique, où ce dernier est plus court (Figure 1.21).

Ce travail a été publié dans *The Plant Journal* en 2009 [Picault et al., 2009].

Route66 *Route66* est un rétrotransposon à LTR de type *Copia*, et qui a été montré comme ayant subi un transfert horizontal de grande envergure entre différentes espèces de riz mais aussi entre différentes *Poaceae* ! Cet élément est un élément récent, complet et *a priori* autonome. Au cours de cette étude, j'ai effectué entre autre une annotation fonctionnelle de l'élément et de ses différentes copies (Figure 1.22).

Ce travail a été publié dans *BMC Evolutionary Biology* en 2009 [Roulin et al., 2009].

Tos17 L'élément le plus connu chez le riz, *Tos17*, est un élément de type *Copia*, actif dans la régénération, la croissance cellulaire, entre les variétés, lors de croisements inter-sous-spécifiques... Lors de son annotation fonctionnelle, je me suis aperçu que *Tos17* ne possédait pas de séquence de GAG (Figure 1.23), protéine obligatoire dans son cycle de rétrotransposition. Après analyse de ses signaux de régulation, il apparait que *Tos17* parasite très probablement *Lullaby*, expliquant la dichotomie entre le fort niveau d'expression et le faible taux d'insertion de *Lullaby*, comme dans le couple *BARE*-1/*BARE*-2.

Cette analyse a donné lieu en 2014 à une publication dans *Mobile DNA* [Sabot, 2014].

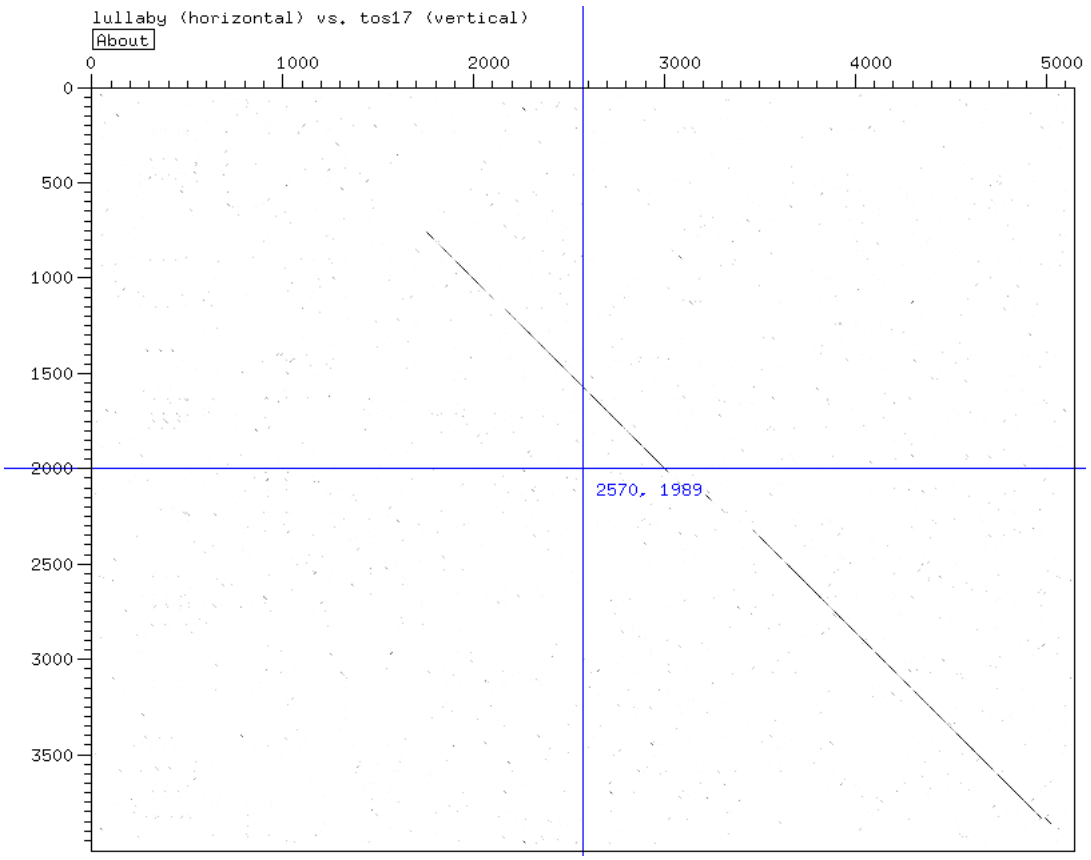

FIGURE 1.21 – *Dotter* de comparaison entre *Lullaby* (horizontal) et *Tos17* (vertical).

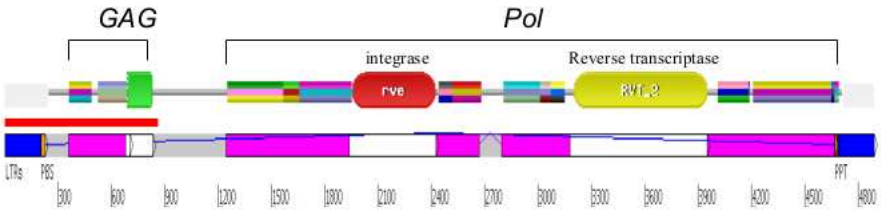

FIGURE 1.22 – Structure de l'élément *Route66*. En bleu foncé sont figurés les LTRs.

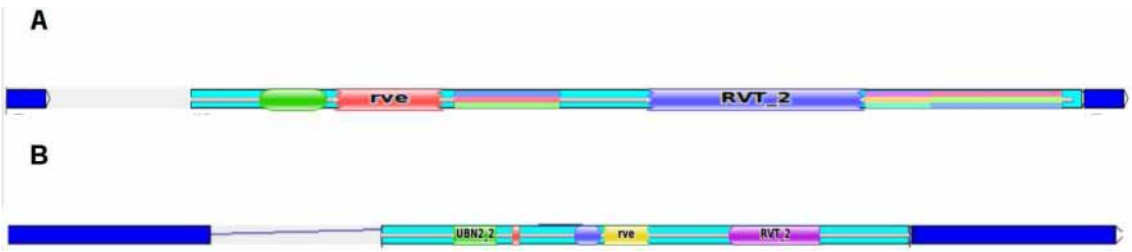

FIGURE 1.23 – Comparaison des structures entre les deux rétrotransposons à LTR de type *Copia*, *RIRE-1* (A) et *Tos17* (B).

1.3.3 Détection des éléments transposables actifs par reséquençage massif

Suite à la difficulté de détecter des éléments transposables transpositionnellement actifs *via* des approches transcriptomiques, je me suis posé courant 2008 la question d'une nouvelle approche pour identifier ces éléments. Il est vite apparu que la meilleure possibilité était non pas d'essayer "d'attraper au vol" les éléments lors de leur transposition, mais bien d'identifier les néoinsertions *après* la

transposition.

Mise au point de la méthode Pour ce faire, nous avons décidé de re-séquencer une plante de Nipponbare après régénération, en choisissant une plante ayant subi plusieurs semaines de culture en cal, et ayant plusieurs néoinsertions de *Tos17*. Cette plante a été séquencée au *Cold Spring harbor Laboratory* en *Illumina* 36 PE, avec une taille moyenne d'insert de 350 bases et une profondeur de 4x.

La méthode employée était basée sur la détection des ruptures de colinéarité (Figure 1.24) entre la séquence de référence et notre plante reséquencée. J'ai utilisé dans le développement de cette méthode différents outils plus ou moins adaptés (*BLAST*, *MAQ*, *Bowtie*...), pour finalement fixer comme outil, avec un étudiant de Master 2 (Moaine Elbaidouri) et un IE bioinformatique (Cristian Chapparo), le logiciel *Bowtie*, pour des raisons de vitesse, même si *MAQ* restait le plus précis. Une analyse post-pipeline a permis de restreindre les faux positifs, et une validation PCR de présence/abscence de confirmer (ou infirmer) les néoinsertions (Figure 1.25).

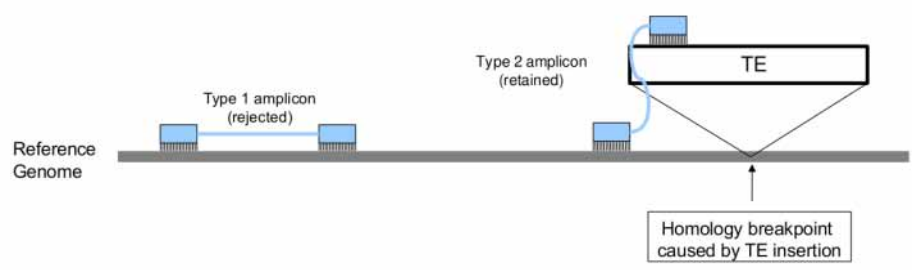

FIGURE 1.24 – Principe d'identification basé sur la rupture de colinéarité entre le génome de référence et le génome séquencé, détectées par les anomalies de mapping des séquences pairées.

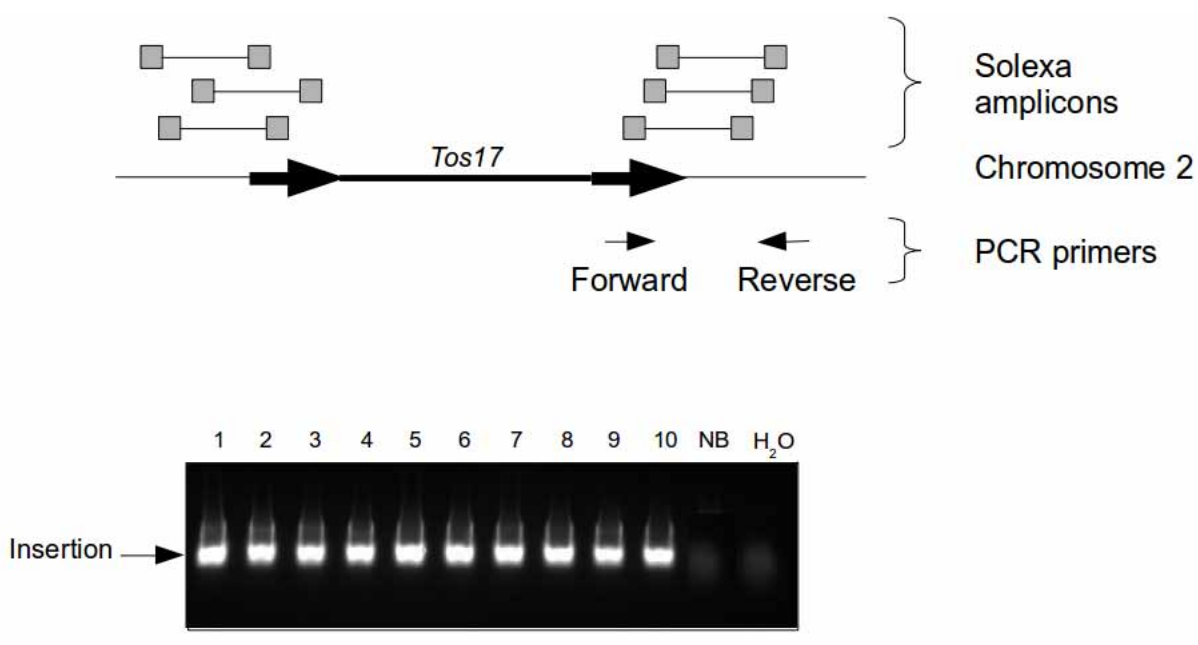

FIGURE 1.25 – Principe de validation PCR entre l'élément cible et la nouvelle localisation d'insertion. **A**) Position des *reads* et des primers PCR, avec l'exemple de *Tos17*. **B**) Résultats de la validation d'une insertion sur 10 descendants de la lignée régénérée (1-10) et sur la variété Nipponbare de référence (NB).

Les limites de la méthode étaient la faible couverture, la courte taille des séquences *Illumina*, et la restriction des sites de cassures de colinéarité à des positions non-ambigües (en général des zones géniques ou simple copie).

Identification des familles connues En plus des candidats classiques tels que *Tos17* et *mPing*, cette analyse a permis à elle seule d'identifier 12 nouvelles familles d'éléments transposables comme actifs (Figure 1.26). Ces éléments sont de Classe I comme de Classe II, avec une forte proportion de MITEs. A noter que la plupart de ces éléments se sont insérés dans ou à proximité de gènes 1.2, sans pour autant affecter le phénotype de la plante en serre en conditions normales. De nombreux éléments n'ont pas été validés du fait de leur insertion dans d'autres éléments, et beaucoup de séquences non identifiées comme des éléments transposables se sont avérées être des sources de rupture de colinéarité.

Des éléments ayant un très grand nombre de copies, comme *RIRE2* (plus de 1000 copies), supposés complètement immobilisés par les mécanismes de contrôle épigénétique, ont eu de nouvelles insertions (Figure 1.26). Enfin, nous avons constaté des variations déjà pré-existantes entre notre Nipponbare de référence utilisé en PCR et la séquence publiée en 2004.

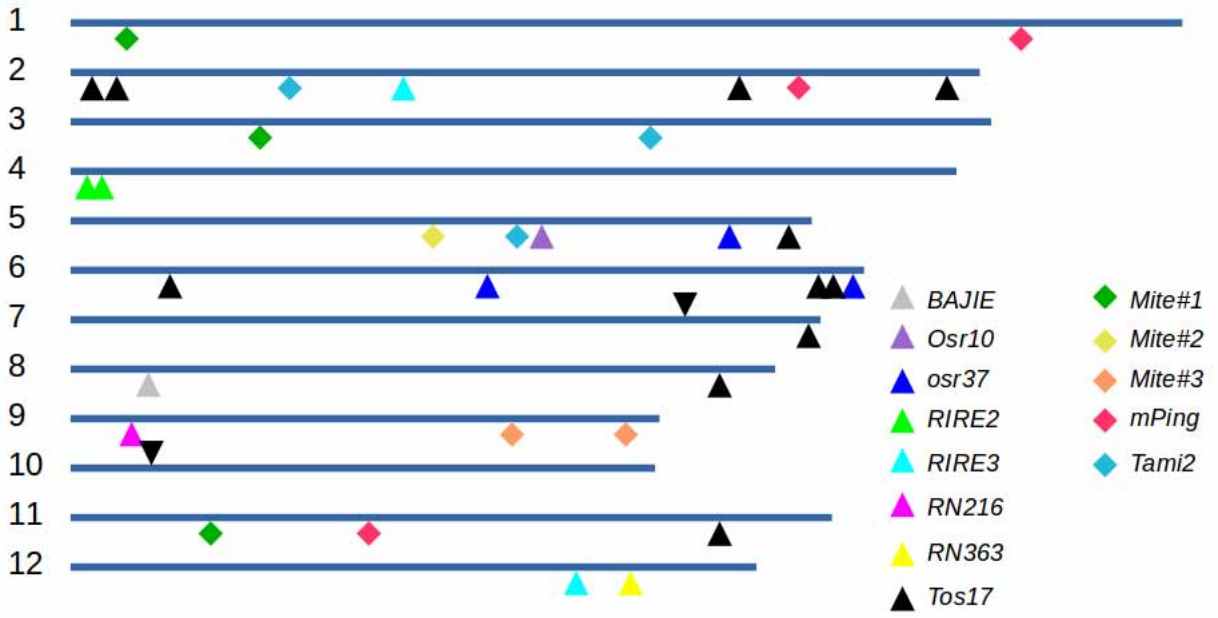

FIGURE 1.26 – Nouvelles insertions détectées dans le cas de la lignée régénérée. En noir sont symbolisées les éléments *Tos17*, en couleur les différentes néocopies des autres familles.

Ce travail a fait l'objet d'une publication en 2011 dans *The Plant Journal* [Sabot et al., 2011].

1.3.4 Apport de la bioinformatique et des NGS à l'étude des éléments transposables

Au cours de cette période, j'ai constaté les possibilités offertes par la génomique massive et plus particulièrement par les approches NGS/bioinformatiques sur l'analyse des éléments transposables. Ces méthodes ouvrent des perspectives très intéressantes dans le domaine de l'écologie génomique, et j'ai donc décidé à cette période de me réorienter complètement sur la bioinformatique pour traiter les données massives de génomique. J'ai d'ailleurs co-écrit un chapitre d'ouvrage avec C. Chapparo sur ce sujet dans une édition spéciale de *Methods Mol Biol* sur les éléments transposables en 2012 [Chaparro and Sabot, 2012].

Ces approches permettent par exemple d'envisager de connaître l'intégralité des nouvelles insertions au cours des générations, en conditions normales ou suite à un stress (comme par exemple la régénération), avoir accès aux insertions somatiques différentielles (d'une feuille à l'autre)... Nous pouvons aussi déterminer plus précisément les moments d'expression et les méthodes de contrôle des ETs, *via* les séquençages RNA et smallRNA.

TABLE 1.2 – Localisation des insertions à proximité ou dans les gènes annotés.

Chromosome	Position	GENE ID	GENE NAME	Relative Position	INTERNAL	TE TYPE
2	1001426	Os02g0118900	NBS-LRR disease resistance protein, putative	0	Exon	*Tos17*
2	1657933	Os02g0131800	Similar to root-specific metal transporter	1000		*Tos17*
2	29511802	Os02g0696500	glycosyl hydrolases family 16, putative, expressed	0	3'UTR	*mPing*
2	35472538	Os02g0809401	Conserved Hypothetical Protein	-400		*Tos17*
4	909838	Os04g0115500	Zinc finger, C2H2-type containing domain protein	-1000		*RIRE-2*
5	14832678	Os05g0319900	Similar to (1.4)-beta-xylan endohydrolase, isoenzyme X-II (EC 3.2.1.8) (Fragment)	0	EXON	*Mite#2*
5	18323796	Os05g0378800	methyltransferase, putative, expressed	0	Exon-Intron	*Tami2*
5	29177343	Os05g0584600	AAA family ATPase, putative, expressed	0	Exon-Intron	*Tos17*
6	31913095	Os06g0728766	Glycosyl transferase, family 48 protein	-900		*osr37*
7	30050157	Os07g0691100	Similar to Pectin methyltransferase 6 (Fragment)	300		*Tos17*
8	3226371	Os08g0155700	Similar to RNA polymerase II largest subunit (Fragment)	0	EXON	*Bajie*
8	26383373	Os08g0528100	Conserved Hypothetical Protein	0	Exon-Intron	*Tos17*
9	18074709	Os09g0460500	gibberellin receptor GID1L2, putative, expressed	200		*Mite#3*
11	5770880	Os11g0211300	Similar to NBS-LRR disease resistance protein homologue (Fragment)	-100		*Mite#1*
11	26064634	Os11g0621300	DUF1399 containing protein, putative, expressed	0	EXON	*Tos17*

1.4 IRD - Équipe GDR/RIcE - Montpellier

En Juillet 2010, du fait des remaniements quadriennaux des UMRs, j'ai quitté le LGDP de Perpignan pour l'équipe GDR (Génome et Développement des Riz) de la future unité DIADE de Montpellier. Je fais toujours partie de cette équipe, qui évolue légèrement dans le prochain plan quinquennal pour devenir l'équipe RIcE (*Rice, Interspecific, Evolution*).

1.4.1 Les Riz Africains

Oryza glaberrima... L'équipe GDR/RIcE travaille principalement sur le modèle du riz Africain cultivé *Oryza glaberrima*. Cette espèce de riz est endémique d'Afrique de l'Ouest, et n'est cultivée que sur cette zone géographique. Cette plante présente de gros intérêts en agriculture africaine, par son potentiel de résistance à des stress biotiques (virus, bactéries, nématodes, adventices...) et abiotiques (sécheresse, toxicité ferreuse...) rencotrés sur le continent Africain [Sarla and Mallikarjuna Swany, 2005]. Malheureusement, *O. glaberrima* est peu productif en comparaison de son cousin asiatique, mais est utilisé comme source de résistance pour *O. sativa*, *via* des croisements interspécifiques. Ces croisements donnent très difficilement des hybrides fertiles, du fait de nombreux facteurs de stérilité interspécifiques, dont le plus connu et le plus fort est le locus S^1 [Sano, 1990, Garavito et al., 2010]. L'AfricaRice a néanmoins obtenu dès le début des années 2000 des lignées issues de tels croisements (lignées NERICA, *NEw RICe for Africa* ; World Food Award 2004) [Gridley et al., 2002], et le Dr Mathias Lorieux a développé depuis de nombreuses ressources agronomiques (*iBridges*) [Garavito et al., 2010] et génétiques (*Chromosome Section Substitution Lines* CSSL) [Gutierrez et al., 2010] pour utiliser *O. glaberrima* pour améliorer *O. sativa*.

... Et son ancêtre *O. barthii* *O. glaberrima* a été domestiqué il y a probablement entre 1000 et 2000 ans dans le delta du fleuve Niger, près de Dia au Mali, depuis l'espèce sauvage *O. barthii* [Portères, 1962]. Cette domestication a conduit à la sélection de traits classiques pour une espèce domestiquée : forte autofécondation, augmentation du rendement, annualité, et perte sensible de l'égrenage. En contrepartie, la variabilité de l'espèce domestiqué est plus faible que celle de l'espèce sauvage, par le phénomène du goulot d'étranglement. Dans le cadre des riz Africains, un autre goulot d'étranglement a eu lieu en amont de la domestication, lors de la spéciation de *O. barthii* depuis l'ancêtre des *O. barthii* /*O. rufipogon* (Figure 1.27).

Le Programme GRiSP Dans le cadre de nos collaborations internationales avec les Centres Internationaux (CIAT [*Centro Interancional de Agricultura Tropical*], IRRI [*International Rice Research institute*] et AfricaRice), nous sommes intervenants dans le programme GRiSP (*Global Rice Science Partnership*), en particulier dans les étapes en amont de la sélection. Le programme GRiSP est le premier des MégaProgrammes du CGIAR (*Consultative Group on International Agricultural Research*, qui dépend de l'UNESCO), regroupant plus de 900 équipes dans le monde, articulés autour de plusieurs Thèmes couvrant l'intégralité de la filière Riz. L'équipe GDR/RIcE est un des piliers du programme, notamment dans la partie de génération de nouveaux matériels à destination des sélectionneurs et d'identification de la variabilité existante dans les riz (Thème 1).Toutes mes recherches depuis 2010 rentrent dans les apports de la recherche fondamentale au programme GRiSP, certaines ayant été financées par le GRiSP (MENERGEP et GLASS).

1.4.2 Transcriptomique & programme Arcad SP1/SP4

Dans le cadre du programme Arcad de la fondation Agropolis, j'ai participé aux Sous-Programmes 1 et 4, "effets de la domestication" et "bioinformatique", respectivement. Il s'agissait de tenter de trouver des signatures de sélection communes à plusieurs espèces tropicales domestiquées, dont le palmier, le caféier, le sorgho, la vigne, l'olivier, et le riz Africain.

Méthodologie et Analyses préliminaires Pour ce faire, nous avons fait séquencer le transcriptome de feuille & d'inflorescence de 10 individus de *O. glaberrima* et 10 *O. barthii*, ainsi que le trans-

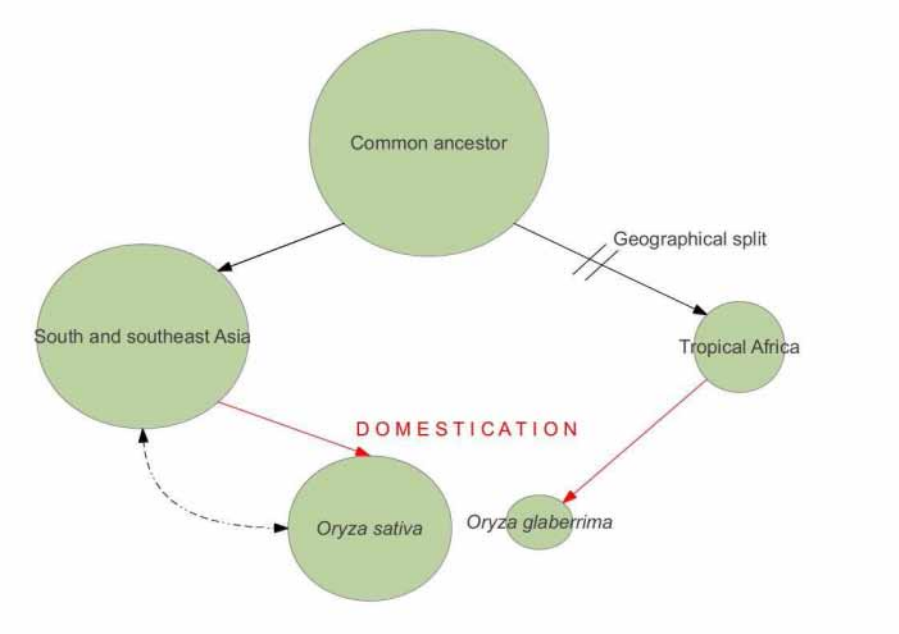

FIGURE 1.27 – Double goulot d'étranglement des Riz Africains. La taille des cercles représente la variabilité génétique de l'espèce (non proportionnelle). Le premier goulot (trait plein noir) a eu lieu lors de l'isolement d'une partie des individus de l'espèce ancêtre en Afrique, par la création de zones refuges lors des dernières glaciations [Vaughan et al., 2008]. Ce goulot a été beaucoup moins fort pour les Riz Asiatiques. Ensuite, la domestication (trait rouge) de *O. glaberrima* a encore réduit le pool génétique et la variabilité du Riz Africain cultivé. Le même phénomène a eu lieu lors de la domestication de *O. sativa*, mais de nombreuses introgressions entre sauvages et cultivés ont eu lieu en Asie, augmentant par-là même la diversité de *O. sativa*. De Cécile Monat.

criptome d'un individu de *O. meridionalis*, une autre espèce du génome AA. Enfin, nous avons utilisé le transcriptome prédit de *O. sativa* Nipponbare comme *outgroup*. Ces séquençages ont été réalisés en *Illumina*, avec des données de 75 à 100 PE, et de 9 millions à 20 millions de pairs de séquences par individu.

Gautier Sarah (IE CIRAD/INRA) et moi-même avons effectué un très gros effort de *benchmarking* à cette occasion, pour tester et identifier les logiciels les plus à même de donner des résultats biologiquement cohérents dans nos analyses, et avons ainsi sélectionné *CutAdapt* [Martin, 2011], *BWA* [Li and Durbin, 2009], *SAMtools* [Li et al., 2009], *PicardTools* et le *GATK* [McKenna et al., 2010] pour les nettoyages, mapping et appels de SNP. L'appel de SNP a été aussi testé avec une "solution maison" de l'équipe de Sylvain Glémin de l'ISEM. J'ai à cette occasion écrit et validé la première version de production des *pipelines* d'analyse Arcad.

Lors de l'application sur *O. glaberrima*, nous avons eu la mauvaise surprise de constater que la séquence de référence du cultivar CG14 fournie par le consortium OMAP [Wang et al., 2014] comportait des gaps très importants non attendus, en plus de fortes restrictions d'utilisation (à cette époque). Nous avons obtenu de meilleurs résultats de mapping en utilisant le génome de l'espèce Asiatique comme référence. Une fois les alignements obtenus et nettoyés, la caractérisation des SNPs a pu avoir lieu, sur plusieurs milliers de gènes pour chacun des 10+10 individus.

***O. glaberrima*, une espèce très peu variable** Le résultat majeur de cette analyse est que l'espèce *O. glaberrima* est à ce jour l'espèce étudié ayant la plus petite base génétique chez les plantes. L'appauvrissement génétique lié à la domestication ("coût de la domestication") correspondrait à une réduction de 20 fois la population de l'espèce ancestrale sur 1000 générations. La base génétique observée est très restreinte, même en incluant des écotypes très différents. Néanmoins, *O. barthii* semble assez divers sur le set de données utilisé.

Les résultats de cette analyse sur une faible partie des riz africains ont donné lieu à une publication

en 2014 dans *Molecular Ecology* [Nabholz et al., 2014].

1.4.3 SNP, MENERGEP & programme Arcad SP2

Toujours dans le cadre du programme Arcad, dans le Sous-Programme 2 "Céréales Africaines" et dans le cadre du projet MENERGEP (programme GRiSP, en collaboration avec AfricaRice), je me suis attelé à une analyse plus large de la variabilité des riz Africains, sur une collection de 280 *O. glaberrima* et de 101 *O. barthii*, en utilisant une approche de type puce SNP *Illumina VeraCode*.

Design de puce SNP à façon Suite à des analyses préliminaires des puces SNPs disponibles en utilisant des ressources génomiques disponibles au laboratoire (2 génomes individuellement séquencés, et deux *bulks* en mélange de 10 *O. glaberrima* et 10 *O. barthii*, respectivement), je me suis rendu compte que les points SNPs utilisés pour le riz Asiatique (puce *44k* de Cornell, [Zhao et al., 2011]) ne seraient absolument pas judicieux pour les riz Africains : seuls 150 points sur 44 000 étaient polymorphes, et fortement liés les uns aux autres. Avec les données de séquençage à ma disposition, et en utilisant un pipeline que j'ai créé et optimisé pour cette analyse, j'ai pu dessiner une puce "à façon" pour les riz Africains, avec un set de 380 SNPs répartis le long du génome (tous les mégabases en moyenne). Nous avons utilisé ces SNPs sur notre collection de riz Africains et un set de 50 *O. sativa* représentatifs de la diversité du riz Asiatique ("mini-core collection" du CIRAD). Au final, seuls deux tiers des points étaient polymorphes (235 points), à cause de faux positifs (de très haute qualité) issus de mutation de type Transversion. Néanmoins, la méthodologie (en prenant en compte les erreurs de type Transversion) est complètement exploitable pour quelque soit l'organisme, à faible coût.

***O. glaberrima* et *O. barthii* sont parmi les espèces les moins variables au monde...** Une fois les différents génotypes obtenus, et avec Mme Julie Orjuela (IE IRD sous contrat MENERGEP), nous avons pu rélaiser une analyse globale de la diversité des riz Africains, en termes de génétique des populations (Figure 1.28).

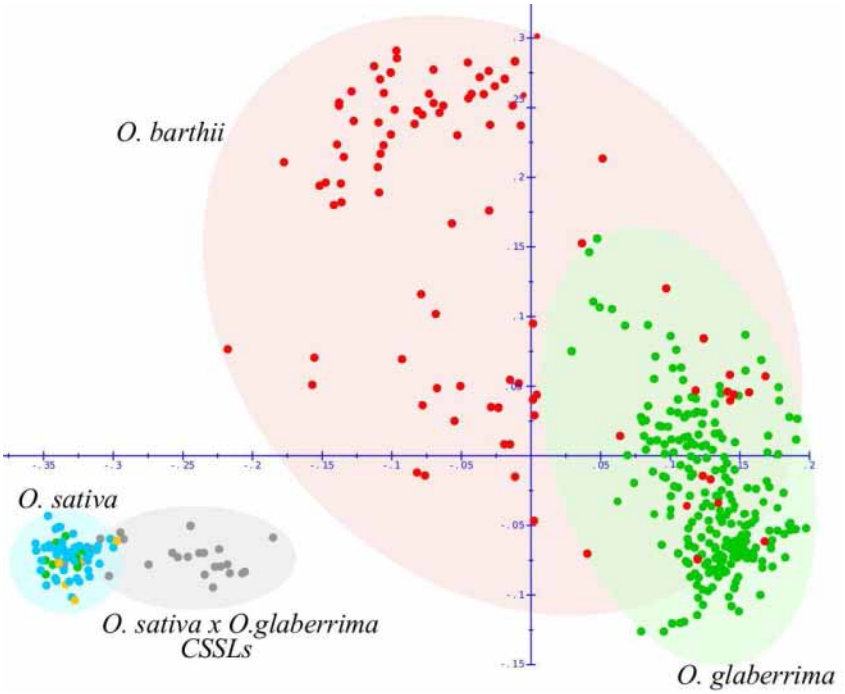

FIGURE 1.28 – Analyse en Composante Principale (Axes 1 et 2) de la variabilité de *O. sativa* en bleu, *O. longistaminata* en jaune, *O. glaberrima* en vert et *O. barthii* en rouge. Les points gris représentent les individus *CSSL* ou des hybrides de laboratoire. Les zones colorées ont été manuellement ajoutées.

Les compartiments Africains et Asiatiques sont bien séparés, et les hybrides de laboratoire se situent bien à l'intersection entre les deux groupes. Contrairement à ce qui a été publié auparavant

[Semon et al., 2005], nous n'avons observé aucun individu naturel présentant un *pattern* de polymorphisme indiquant un hybride entre *O. glaberrima* et *O. sativa*. Les hybrides putatifs décrits par Semon et al. correspondraient plus à une forme différente de *O. sativa*.

Les deux espèces Africaines sont bien séparées (Figure 1.29), avec un F_{st} de 0.212. La zone de recouvrement des deux espèces sur l'ACP correspond à des individus adventices classés auparavant comme *O. barthii*, et issus de croisements interspécifiques entre le sauvage et le cultivé.

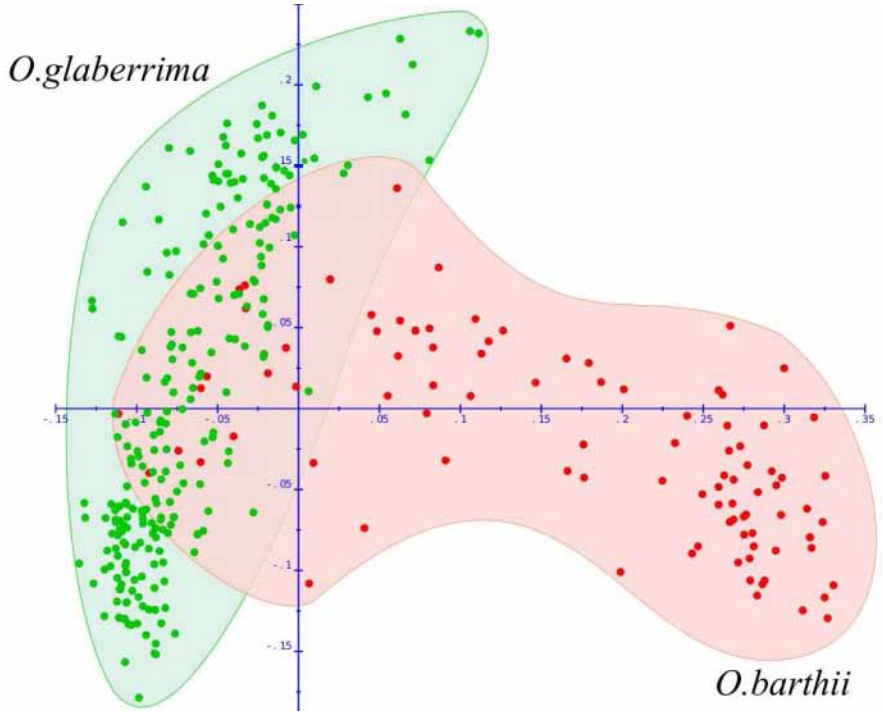

FIGURE 1.29 – Analyse en Composante Principale (Axes 1 et 2) de la variabilité de *O. glaberrima* en vert et *O. barthii* en rouge. Les zones colorées ont été manuellement ajoutées.

Le riz Africain sauvage *O. barthii* est comme attendu plus variable que le cultivé, mais cela reste quand même très faible. *O. barthii* a pu être structuré en trois populations (Figure 1.30) liées à la géographie, avec la population 1 couvrant l'Afrique Centrale, la population 2 la région du lac Tchad, et la population 3 l'Afrique Australe et Occidentale, incluant le Delta du fleuve Niger. Un certain nombre d'admixtures existe entre ces trois populations, indiquant un flux migratoire constant (valeurs de F_{st} entre 0.244 et 0.311).

Pour *O. glaberrima*, nous avons pu extraire deux populations (Figure 1.31). Comme attendu pour cette espèce la variabilité est très faible, beaucoup plus que pour *O. barthii*, avec un très fort inbreeding (F_{is} de 0.970 et 0.932). Aucunes liaisons entre des phénotypes ou des écotypes connus n'ont pu être révélés dans ces deux sous-populations, ni aucune structuration géographique. De plus, un grand nombre d'admixtures (53) existent entre ces deux groupes.

En conclusion, c'est la première fois qu'une structuration génétique a pu être identifiée pour les *O. barthii*, et la position de plusieurs individus considérés comme hybride a pu être clarifié. De plus, la population ainsi génotypé sert de matériel de base à nos analyses actuelles et futures (Programme IRIGIN, voir plus loin). Le set de SNPs utilisé a été transféré pour à nos partenaires d'AfricaRice et du *Generation Challenge Program* pour être utilisé dans le cadre des programmes de Ressources Génétiques et d'amélioration variétale des CG.

Ce travail a donné lieu à une publication en 2014 à *TAG Theoretical & Applied Genetics* [Orjuela et al., 2014].

1.4.4 smallRNA & Riz Africains

Les petits ARNs Les petits ARN (smallRNA) sont, comme leur nom l'indique, des petits fragments d'ARN simple brin de 18 à 28 bases, impliqués dans la régulation de l'expression des gènes, dans le

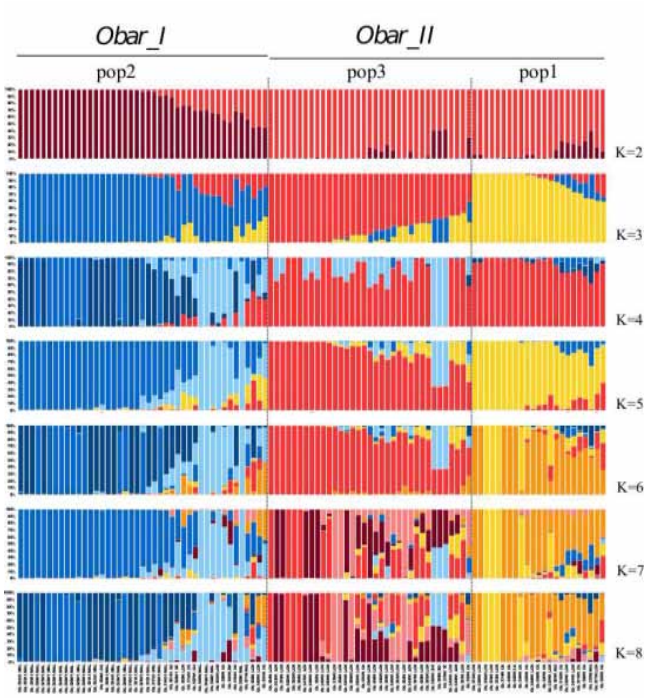

FIGURE 1.30 – Analyse *STRUCTURE* de 98 *O. barthii*. Les meilleures statistiques sont obtenues pour K=2 et 3.

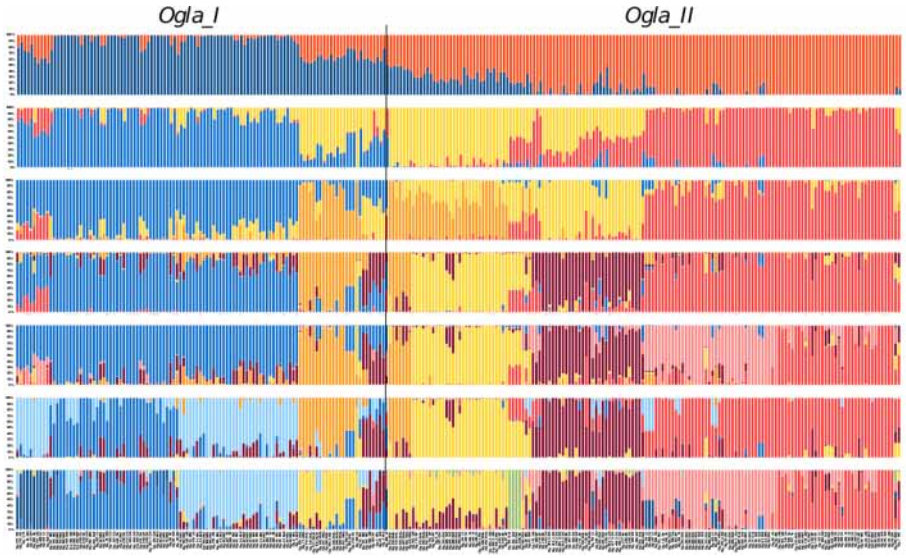

FIGURE 1.31 – Analyse *STRUCTURE* de 260 *O. glaberrima*. Les meilleures statistiques sont obtenues pour K=2.

contrôle des éléments transposables et dans la mise en place de la méthylation [Slotkin and Martienssen, 2007]. Dans leur action de régulation des gènes, ce sont surtout les microRNAs (miRNAs), de 20-22 bases, qui sont actifs, *via* une action post-transcriptionnelle couplée avec les protéines des groupes *Dicer* et *Argonaute*. Cette action ciblée (reconnaissance de séquence miRNA/ARNm) provoque une dégradation de l'ARNm cible. Dans différentes études sur le développement, il a été montré que c'est l'action des miRNAs, et non pas une expression différentielle des ARNm, qui est à l'origine de certaines des différentiations cellulaires et tissulaires [Schmitz and Ecker, 2012]. Dans le cadre d'une étude comparée entre *O. glaberrima* et *O. barthii* sur le développement paniculaire, nous avons réalisé une analyse en *bulk* des petits ARNs correspondants aux premiers stades de différentiations tissulaires du méristème

inflorescentiel. L'idée initiale était d'identifier des miRNAs différentiellement exprimés entre le sauvage et le cultivé, à mettre en lien avec les différences de niveau de branchement secondaires et tertiaires.

Méthodologie & Implémentation Ces petits ARNs ont été extraits de 10 plantes de chacune des deux espèces, aux mêmes stades, puis séquencés en *Illumina* 50 SE. Une fois les séquences nettoyées des adaptateurs et des bases de mauvaises qualité, nous avons "clusterisé" les séquences identiques, puis les avons traitées comme de simples données FASTA.

J'ai ensuite développé un pipeline qui permet d'assigner à chaque séquence une cible potentielle, *via* une analyse hiérarchique utilisant *BLAST*. En fixant un seuil de taille et d'identité minimaux, il m'a été possible de classer la plus large majorité des séquences, suivant la hiérarchie suivante : *microRNAs*, *structural RNAs*, *Transposable Elements*, *Genes*/CDS, *Genes*/introns-UTRs, *unannotated regions*, *unmapped regions* (Figure 1.32).

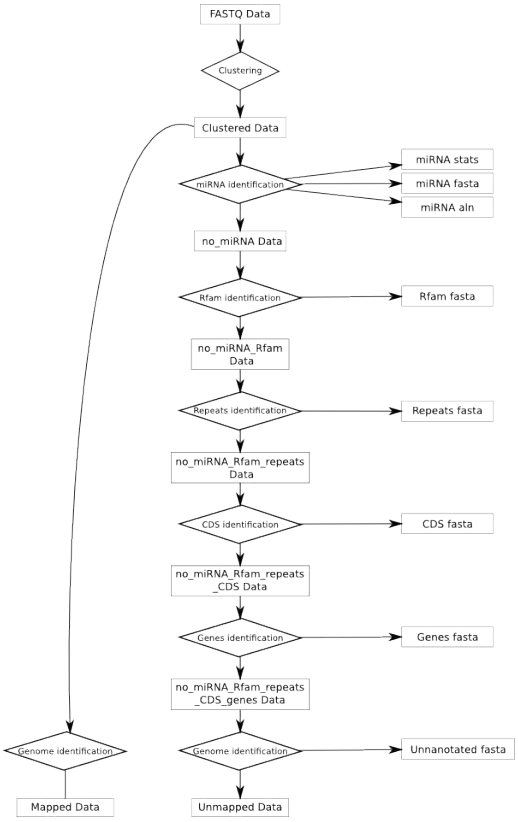

FIGURE 1.32 – Pipeline d'analyse *BLAST* des petits ARNs

En analysant plus finement les données, nous nous sommes aperçus qu'une forte proportion (près de 45%) des petits ARNs de 21 nucléotides étaient plus exprimés chez *O. barthii* que chez *O. glaberrima* (Figure 1.33).

Ces séquences sur-exprimés chez les 21nt semblent de plus être liées à des zones non-annotées du génome (Figure 1.34).

Les phasiRNA, marqueurs moléculaires des stades de développement du méristème inflorescentiel Toujours en allant plus loin, nous nous sommes aperçu que certains miRNAs étaient eux aussi sur-exprimés, dont le miR2118. Ce miRNA est impliqué dans la formation secondaires de petits ARNs de 21nt dits phasés (Figure 1.35), qui sont issus d'un ARN précurseur plus long (120 bases minimum) [Johnson et al., 2009, Fei et al., 2013]. Ces phasiRNAs avaient déjà été identifiés chez le riz Asiatique *O. sativa* dans des panicules, sans qu'aucun rôle putatif ne leur ai été attibué.

Ici, par comparaison entre le sauvage et le cultivé, nous avons identifié une liaison entre la quantité de ces petits ARNs phasés et le niveau de branchements secondaires des panicules de riz, *via* la vitesse de différentiation des méristèmes inflorescentiels. En effet, d'autres analyses ont montré (Thèse de Ta

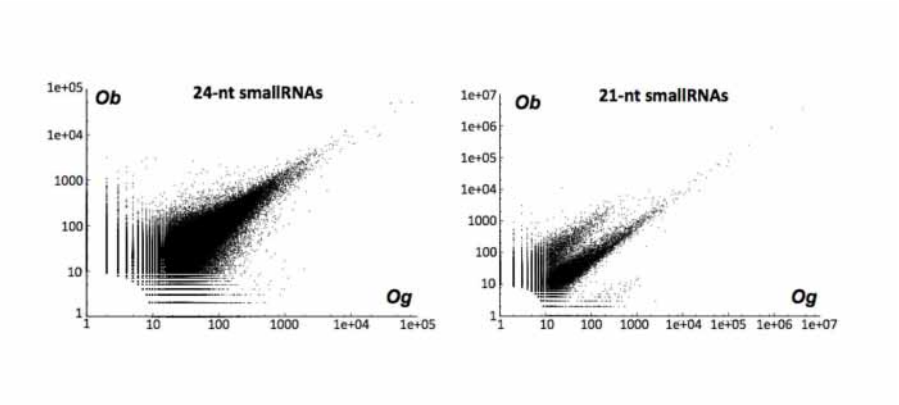

FIGURE 1.33 – Quantité relative des petits ARNs 21/24 nucléotides entre *O. barthii* (Ob) et *O. glaberrima* (Og)

.

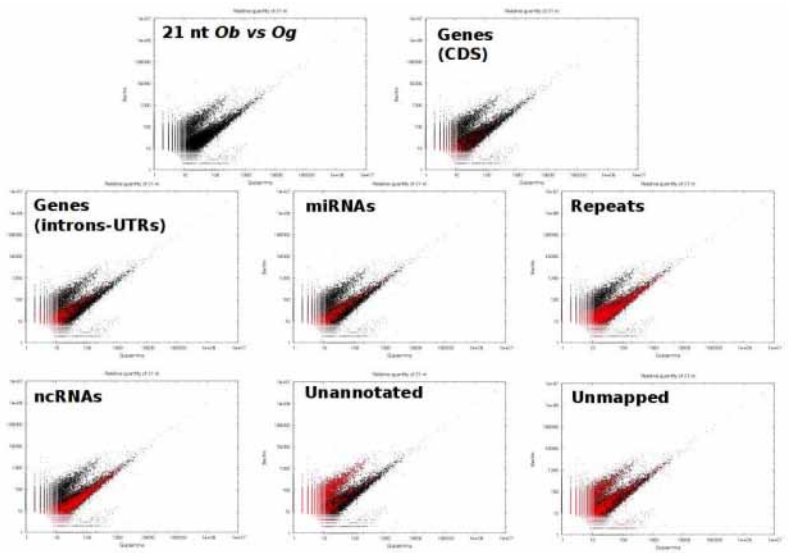

FIGURE 1.34 – Quantité relative des petits ARNs 21/24 nucléotides entre *O. barthii* (Ob) et *O. glaberrima* (Og) en fonction des différents compartiments génomiques. En noir sont notés l'intégralité des points, en rouge les points issus de la fraction considérés.

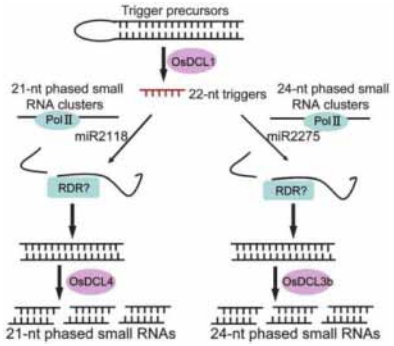

FIGURE 1.35 – Mécanisme d'action des miR2118 et 2275 sur les précurseurs de phasiRNA. De [Johnson et al., 2009]

.

Kim Nhung) que des gènes marqueurs de différentiation étaient à un stade plus avancé chez le sauvage que chez le cultivé, et ce avec les mêmes stades histologiques apparents. Les ARNs phasés sont donc liés à cette différentiation, mais nous ne savons pas encore de quelle manière : sont-ils les initiateurs ou les résultats de cette différentiation ? Pour le moment aucune de nos données ne permet de trancher, et ces ARNs phasés restent donc pour le moment un bon marqueur du stade réel de différentiation.

Ce travail a été soumis pour publication en 2014 à *Plant Physiol.*

1.4.5 Assemblage des génomes des Riz Africains - Le projet GLASS

Pourquoi refaire un assemblage ?

Au cours des analyses des données transcriptomiques de *O. glaberrima*, comme indiqué avant, nous nous sommes aperçu que la séquence fournie par le consortium OMAP du cultivar CG14 était incomplète [Wang et al., 2014]. Pire encore, elle était moins efficace en termes de mapping que la référence de *O. sativa*, Nipponbare pour les analyses des données transcriptomiques de CG14 lui-même ! En effet, seul 70% des données de séquences étaient mappées sur OMAPv1, alors que ce taux montait à 83% sur Nipponbare dans les mêmes conditions. De plus, l'analyse des données génomiques lors de la fabrication de la puce riz Africains a montré que plusieurs grandes zones étaient dupliquées dans le génome Africain par rapport à l'Asiatique. Enfin, le cultivar CG14 ne représente qu'un type d'écotype de *O. glaberrima*, les riz pluviaux *upland*.

Suite à ces constatations, nous avons décidé, en collaboration avec AfricaRice, de générer nos propres références pour *O. glaberrima*, et ce sur deux cultivars emblématiques, CG14 et TOG5681 : c'est le projet **GLASS** (pour *GLaberrima ASSembly*), dont j'ai été le porteur. Ce projet courrait sur l'année 2013, et a impliqué une étudiante de Master2 (Cécile Monat, actuellement en Thèse de Doctorat sous ma direction).

Méthodologie & Résultats

Séquençage Nous avions choisi initialement une approche hybride *Illumina/PacificBiosciences* pour générer nos assemblages. Pour ce faire, nous avons obtenu pour chacun des deux individus une couverture de 60x en *Illumina* (100 PE) et de 11x en *PacBio* (RS2, taille moyenne de 5kb, jusqu'à 28kb). Nous comptions utiliser la méthodologie présentée dans [Koren et al., 2012], *via* l'utilisation du logiciel *PacBioToCA*.

Malheureusement, la méthode n'est pas encore au point pour être exportable hors du laboratoire des créateurs des logiciels, et nous n'avons pu obtenir de résultats via cette approche. Nous nous sommes donc rabattu sur une méthode plus classique, en assemblant *via* le logiciel *Abyss* les données *Illumina* dans une premier temps (assemblage v1), puis en utilisant les données *PacBio* avec une approche de *gap-filling* (assemblage v2 en cours).

Résultats de la v1 Les assemblages v1 de TOG5681 et CG14 avec les données *Illumina* ont donné de très bons résultats en utilisant *Abyss*, comme montré sur le Tableau 1.3, avec une couverture de près de 75% équivalent génome (par rapport à *O. sativa*).

Les données sont en libre accès, avec une annotation automatique portée depuis *O. sativa ssp japonica* cv Nipponbare, avec l'aide d'une IE bioinformatique (Christine Tranchant). Un site web a été créé, `http://oryza.mpl.ird.fr`, développé en collaboration avec AfricaRice, pour héberger un *Gbrowse* et un dépôt de données.

Une publication sur ce sujet est en cours, avec une première soumission prévue en Décembre 2014 à *BMC Genomics*.

Prévisions pour la v2 Pour la v2, nous sommes en train d'appliquer une méthode de *gap-filling*, *i.e.* utiliser les données *PacBio* de grande taille pour combler les trous entre nos scaffolds, avec des outils de type *PBJelly*. La majorité des gaps étant formé par des séquences répétées de grande taille non assemblables à partir des données *Illumina* (petites séquences), les données *PacBio*, de plus

	CG14-scaffolds	TOG5681-scaffolds
# sequences after sort	64.988 scaffolds	51.262 scaffolds
N50 after sort	10.233 bp	13.404 bp
N90 after sort	2.025 bp	2.827 bp
# bases after sort	299.704.894 bases	305.237.265 bases
coverage after sort	0,75	0,76
avr lgth of sequences after sort	4.611,7 bp	5.954,45 bp
median length of sequences after sort	2.077 bp	2.733 bp
maxi length of sequences after sort	90.835 bp	105.329 bp

TABLE 1.3 – Statistiques d'assemblage des données *Illumina*. De Cécile Monat

grande taille, permettront de passer outre et de "fermer" les gaps. Nous prévoyons une couverture finale proche des 90% de la séquence complète de *O. glaberrima*.

Toutes ces données nous donneront accès à deux séquences de référence de bonne qualité, pour pouvoir envisager de travailler à l'échelle de l'espèce entière (voir ci-après pour les aspects pan-génomiques).

1.4.6 Mise en place d'un pipeline bioinformatique pour une étude de phylogéographie de quatres espèces africaines par analyse des variations chloroplastiques

Dans le cadre du projet *Chlorodiv*, qui traite de la diversité et de la phylogéographie de 4 espèces Africaines (Fonio, Mil, Igname et Palmier rotin), j'ai participé à la validation bioinformatique de l'assemblage de génomes chloroplastiques à partir de données *Illumina* d'extraits d'ADN total, ainsi qu'à la création d'un pipeline d'identification de SNPs dans différentes populations de ces 4 espèces. Une validation des assemblages réalisés par *MITObim* a été réalisée, et j'ai prototypé la validation et encadré l'étudiant de Master2 bioinformatique (Ayite Kougbeadjo) qui a réalisé l'analyse. Nous avons utilisé des données de séquençage de RAM63, un *O. sativa ssp indica* de type *aus* pour valider la méthodologie (Figure 1.36).

Cette validation des assemblages a été publiée en 2014 dans *Molecular Ecology Ressources* [Mariac et al., 2014].

Le pipeline, développé par le même étudiant, est construit à partir des modules développés dans le cadre du *Dojo code*. Il a permis dès sa première version de multiplier par 4 le nombre de marqueurs disponibles pour la phylogéographie du Mil basée sur le chloroplaste. L'utilisation en a été simplifié au maximum pour les biologistes, de manière à se concentrer sur les paramètres biologiques plutôt que ceux informatiques.

1.4.7 Calcul au Sud et *Single-Board Computer* : le Projet Framboisine

L'objectif du projet Spirale **Framboisine** est de fournir une méthodologie d'assemblage et d'utilisation de mini-clusters de calcul formés de *SingleBoard Computers*. Ces ordinateurs ultracompacts (format carte de crédit, Figure 1.37), de faible coût (de 35 à 100 euros) et de faible consommation énergétique (de 2 à 5 W), sont basés sur des processeurs ARM de diverses générations (processeurs de téléphones portables), et avec des capacités de calcul pouvant aller aux QuadriCoeurs avec 2 Go de RAM. Il est possible de les mettre en cluster, *i.e.* de les monter en structure de calcul parallèle, et ainsi de pouvoir obtenir des machines de calcul certes à performance modérée, mais à très bas coût (moins de 2000 euros par cluster de 20 cœurs et 20 Go de RAM).

De telles machines pourront servir à la fois :

— dans le cadre de formation au Sud (pré-doctorale et doctorale). En effet, une des limites de ces formations est souvent l'accès à des ressources de calcul suffisantes. Il existe des plateformes de calcul distantes (exemple du cluster de calcul bioinformatique de l'IRD) mais l'accès à une connexion réseau stable est parfois problématique. Une machine de ce type coûtant moins de 2000 euros, il sera possible d'en emmener une à chaque fois qu'une école thématique ou une formation au Sud portant sur du calcul parallèle sera effectuée (génomique, épidémiologie, simulation).

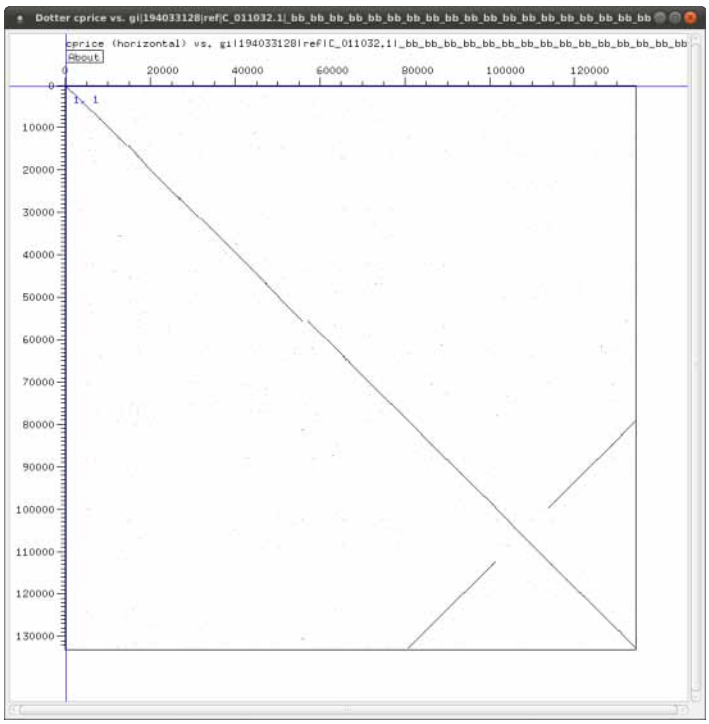

FIGURE 1.36 – *Dot Plot* du chloroplaste reconstruit de RAM63 (vertical) *vs* le chloroplaste de référence des *indica* (horizontal). La zone répétée est bien visible. Les cassures sont dûes à des zones non-présentes dans le génome chloroplastique qui a servi pour reconstruire celui de RAM63 (*Brachypodium dystachion*). De Mariac et al, 2014.

FIGURE 1.37 – *RaspberryPi* utilisé comme noeud dans le montage de certains clusters Framboisine.

Ceci permettra de travailler directement sur la machine et de pouvoir la laisser sur place au bénéfice du partenaire ayant co-organisé la formation.

— dans le cadre de projets de recherche réalisé par nos partenaires localement. Ceci permettra à n'importe quel partenaire du Sud de pouvoir travailler sur des approches demandant du calcul parallèle (génomique, écologie, épidémiologie, simulation météo, astronomie et radioastronomie,...) de manière autonome.

Actuellement nous avons un prototype fonctionnel basé sur le *Raspberry Pi* (Figure 1.37) en asso-

ciation avec le *CubieTruck* (Figure 1.38).

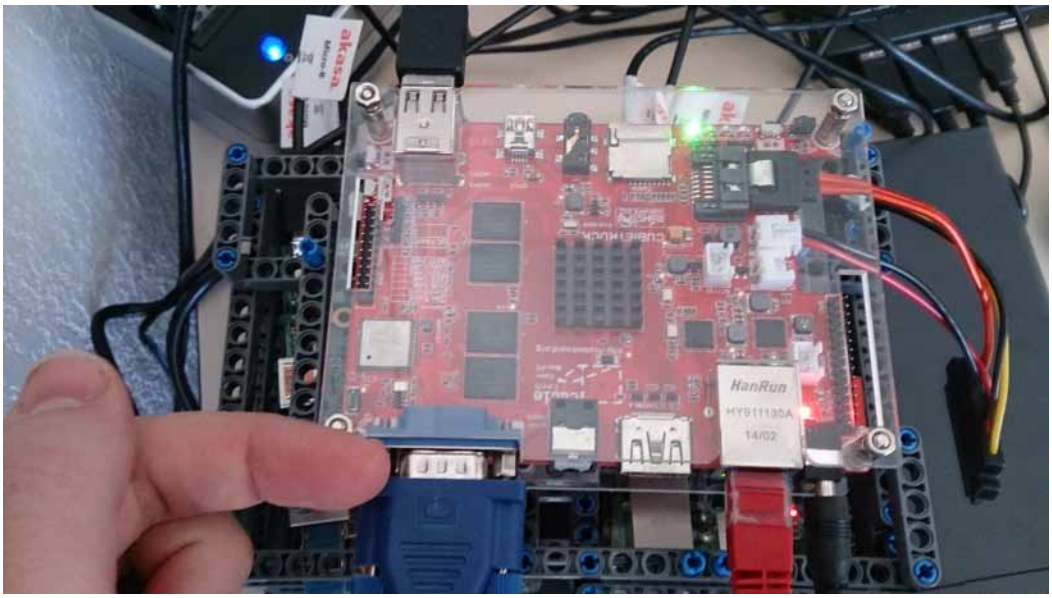

FIGURE 1.38 – *CubieTruck* utilisé comme *Master* dans le montage des clusters Framboisine.

Nous allons tester deux nouveaux modèles de cluster, avec une version entièrement faite de *CubieTruck* et une autre faite à partir de *OlinuXino-A20*. L'assemblage final sera réalisé en utilisant des Legos *Technic* (Figure 1.39), seul matériel convenant à une telle utilisation (pas d'outils, pas de découpe, résistance à la chaleur, à l'humidité, pas de conduction électrique, disponibilité mondiale).

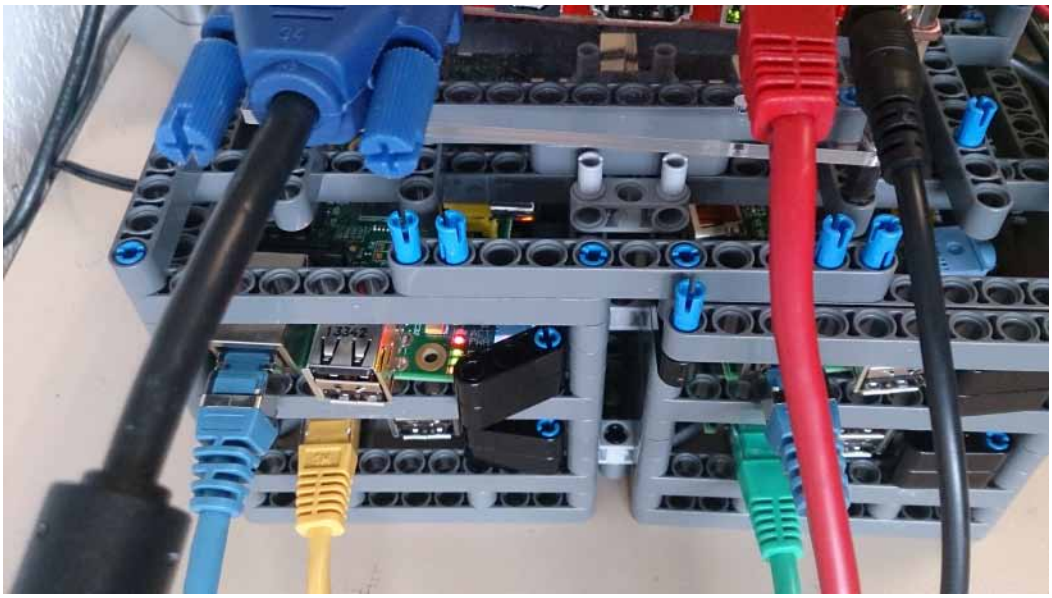

FIGURE 1.39 – Legos *Technic* utilisé dans le montage de certains Framboisine.

L'avancée des travaux est régulièrement présentée sur le facebook du projet. Le cluster *RaspberryPi* a été utilisé avec succès en enseignement en France, au Sénégal et au Vietnam (Figure 1.40). On peut l'emmener dans ses bagages, et ainsi s'affranchir de tout réseau local (ou absence de réseau) dans l'enseignement en bioinformatique.

Trois sites au Sud ont été choisi pour la phase finale de validation en mode recherche, Bel-Air au Sénégal, AfricaRice et FAST/DASSA au Bénin.

FIGURE 1.40 – Utilisation d'un cluster Framboisine en enseignement, ici à l'USTH de Hanoï, VietNam.

1.5 Apport des NGS au concept d'espèce...

Les NGS permettent à l'heure actuelle d'obtenir très facilement l'information génomique pour des dizaines d'individus en même temps, à un coût moindre qu'il y même seulement 3 ans. Même si cette information génomique reste incomplète (du fait des difficultés d'assembler les répétitions par exemple), nous disposons d'un énorme réservoir de données qui nous permet d'aborder sur un nouvel angle une notion fondamentale en biologie : *Qu'est-ce qu'un espèce végétale ?*

En effet, la définition de base de l'espèce, avec les aspects de fertilité, de descendants fertiles, de localisation géographique commune et autres, ne permet souvent pas de définir exactement les espèces chez les plantes. Par exemple, les blés se croisent très facilement entre eux, avec des descendants fertiles, et sont souvent issus de la même zone géographique. De même, *O. barthii* et *O. glaberrima* sont interfertiles et sympatriques...

L'accès aux ressources génomiques des individus de la même espèce va ainsi permettre de déterminer des limites génétiques et génomiques pour ces espèces. Dans ce cadre là entrent en jeu les notions de Pan- et Core-génomes.

1.6 Autres collaborations et projets

En plus de mes projets de recherches, j'ai aussi collaboré de manière plus anecdotique à des projets portés par d'autres chercheurs et équipes,

1.6.1 Annotation automatique des éléments transposables par compte de *k*-mers

Dans le cadre de la mise au point de nouvelles approches d'annotation automatique de éléments transposables, j'ai participé à une courte étude sur l'utilité des k-mers dans cette optique. L'idée était la suivante ; en disposant d'un sous-set de séquence d'un génome donnée (ici 0.1x en *Illumina*), on compte les k-mers d'une taille donnée présent dans ce sous-set. Chaque k-mer possède donc ensuite une valeur numérique. En reportant sur le génome cette valeur à chaque position correspondante, nous obtenons une quantification assez directe de la répétabilité d'une séquence donnée. Ainsi, si une séquence est présente 10 fois dans un génome (e.g. 10 copies d'un élément), les k-mers associés à cette séquence auront tous en moyenne une valeur de 1 dans un sous-set de 0.1x. Ainsi, la courbe obtenue reflètera la densité de copie (Figure 1.41).

Cette courbe de densité a été comparé à une annotation manuelle de 10 grandes séquences d'orge et

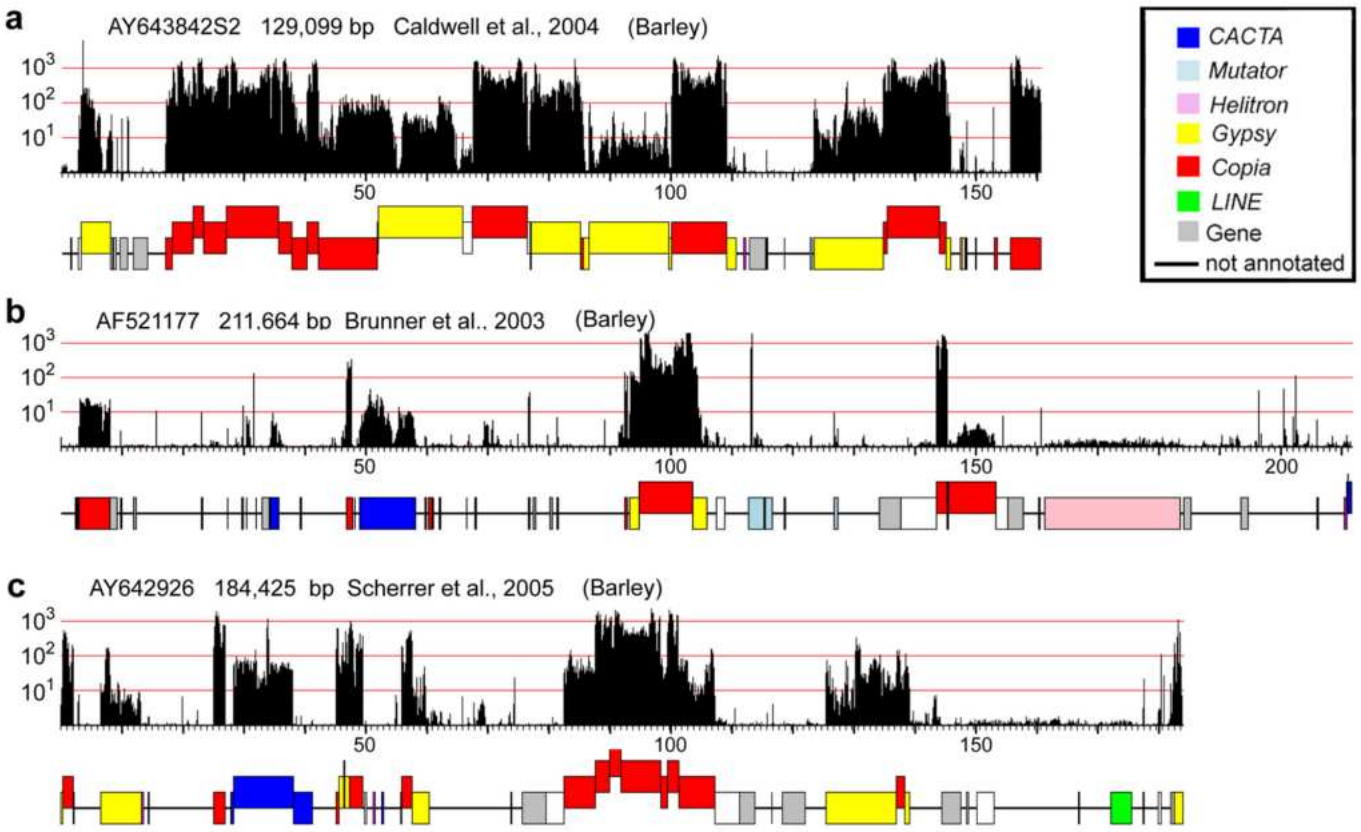

FIGURE 1.41 – La courbe de densité en noir est superposée à l'annotation manuelle des éléments sur 3 régions du génome de l'orge. De [Wicker et al., 2008a].

de 8 de blé (principalement des BACs). En moyenne, 64% des séquences ont été annotées manuellement pour leur contenu en éléments transposables, et 50% *via* le compte des k-mers. Le recouvrement entre les deux méthodes dépassent les 80%, avec des zones identifiées comme ETs uniquement en manuel ou uniquement en k-mers (puis validés manuellement après), facilitant ainsi l'annotation de nouvelles régions. Dans le cadre de cette étude, j'ai effectué l'annotation des séquences de blé, ainsi que la validation des annotations k-mers *a posteriori*.

Ce travail a été publié en 2008 dans *BMC Genomics* [Wicker et al., 2008a].

1.6.2 Annotation des éléments transposables dans le génome du Cacaoyer

J'ai participé au consortium d'annotation du génome du Cacaoyer, où j'ai été en charge des éléments transposables de Classe II de type TIR. J'ai à cette occasion développé une approche itérative basé sur du *fishing* de transposase puis de l'extension de séquence à la recherche de répétitions inversées pour border les éléments, et enfin un nettoyage manuel.

Ce travail a été publié en 2010 dans *Nature Genetics* [Argout et al., 2010].

1.6.3 Diversité au sein du genre *Oryza* de l'élément *Tos17*

L'élément *Tos17* a beau être l'élement le plus actif (*a priori*) chez le riz Asiatique, il n'est pas présent partout dans le genre *Oryza*. Une analyse poussée en biologie moléculaire sur plusieurs individus par espèce du génome AA a en effet montré que *O. glaberrima* en est dépourvu, et que le nombre de copies intra-espèces était très variables (entre 1 et 11 copies chez le riz Asiatique par exemple).

Ce travail a été publié en 2009 dans *Mol Genet Genomics* [Petit et al., 2009].

1.6.4 Éléments Transposables de la vigne

J'ai apporté mon expertise dans le cadre de la thèse de Grégory Carrier sur les éléments transposables de la vigne et les polymorphismes somatiques des clones de vigne.

Cette collaboration a abouti à une publication dans *PLoS One* en 2012 [Carrier et al., 2012].

1.6.5 *MADS box* et anomalie *Mantled* du palmier à huile

L'anomalie *Mantled* provient d'anomalies épigénétiques chez certains plants de palmier à huile issus de culture *in vitro*. Un des gènes à l'origine de cette variation, *EgDEF1*, est une *MADS-box*, classe de gènes responsable entre autre de la floraison. Ici, il a été montré que ce n'est pas une possible méthylation différentielle, ou même une surméthylation/déméthylation d'éléments transposables proches, qui est à l'origine de ctte variation *Mantled*, mais une variation dans le ratio d'expression entre la forme longue et courte du gène *EgDEF1*, ouvrant la voie vers d'autres types de régulation que la simple variation de méthylation.

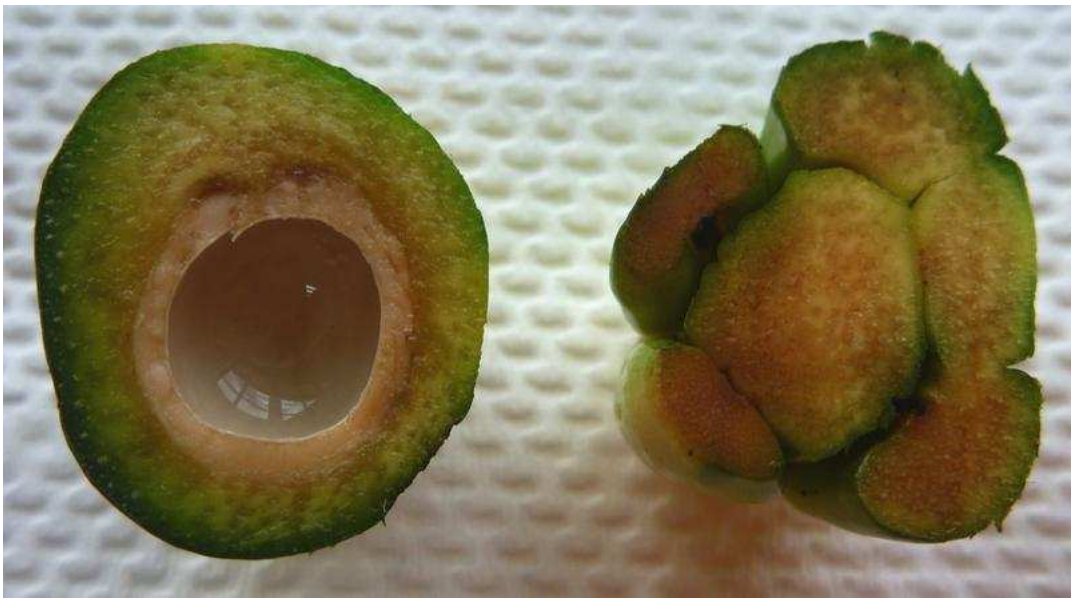

FIGURE 1.42 – Anomalie *Mantled* du palmier à huile. A gauche un fruit normal, à gauche un fruit anormal issu de multiplication *in vitro*. De E. Jaligot.

Ce travail a été publié en 2014 dans *PLoS One* [Jaligot et al., 2014].

1.6.6 IBC, Institut de Biologie Computationelle

L'IBC est un *Investissement d'Avenir en Bioinformatique 2011*, avec comme objectif de mettre en relation des biologistes et des informaticiens pour obtenir une masse critique en recherche en bioinformatique. Cinq axes thématiques sont développés dans l'IBC :

- ***WP1-HTS :*** *Methods for high-throughput sequencing analysis*
- ***WP2-Evolution :*** *Scaling-up evolutionary analyses*
- ***WP3-Annotation :*** *Structural and functional annotation of proteomes*
- ***WP4-Imaging :*** *Integrating cell and tissue imaging with Omics data*
- ***WP5-Databases :*** *Biological data and knowledge integration*

Je travaille au niveau du WP5 comme fournisseur de données et question biologiques, en collaboration avec Pierre Larmande (IE IRD), Manuel Ruiz (Chercheur Cirad) et Patrice Valduriez (DR INRIA). Nos questions tournent autour de la mise en place de *pipelines* dédiés à la mise en relation

phénome/génome, la manipulation des données massives, les bases de données *NoSQL*, le web sémantique et les langages contrôlés... Plusieurs prototypes de bases de données à haute vitesse pour les NGS ainsi que des *pipelines* basés sur *Galaxy* et *OpenAlea* ont déjà été testé.

Chapitre 2

Projet de Recherches

2.1 Différents Projets de Recherches, un même But

J'entreprend actuellement 3 projets de recherche, différents au premier abord, mais qui portent tous sur la même idée, **une meilleure connaissance des génomes des riz Africains cultivés et sauvages**. Ces trois projets vont du plus fondamental, avec l'analyse des Pan et *Core*-génomes, au plus appliqué, avec la mise au point de capture de séquence génique pour séquençage ciblé, en passant par une partie plus en aval de l'agronomie, avec les riz adventices.

2.1.1 Mieux connaître le riz Africain pour extrapoler au riz Asiatique

Le riz Africain est pour nous à la fois un modèle et un objet d'étude. Objet d'étude car il intervient dans beaucoup de variétés agronomiques pour nos partenaires du Sud, entre autres les variétés NERICA, mais aussi pour ses avantages dans les conditions Africaines : résistances aux adventices, à la toxicité ferreuse, à des virus et bactéries, à la sécheresse... De plus, l'espèce domestique *glaberrima* est très peu variable, et son ancêtre sauvage *barthii* l'est à peine plus. Ces deux espèces sont de bons modèles pour identifier les variations dues à la domestication et l'amélioration variétale, dans un contexte beaucoup plus simple que pour le riz Asiatique, dont l'histoire est plus complexe...

Ainsi, même si mes projets portent sur le riz Africain, le point de mire dans le futur lointain est d'utiliser les connaissances et les données issues des riz Africains pour avancer sur le riz Asiatique.

2.1.2 Le Programme IRIGIN, fournisseur de données massives

Pour chacun de ces projets, je m'appuyerai beaucoup sur les données issues du Projet IRIGIN (*International RIce Genomic INitiative*) que je coordonne entièrement, de l'appel d'offre *France Génomique* 2011. Ce projet entre l'IRD, le CIRAD, le CIAT et AfricaRice prévoit le séquençage de milliers de lignées de riz, pour un volume total équivalent à 17 000 fois le génome du riz, incluant 4 sous-projets :

— **Séquençage des Riz Africains** : 400 individus sélectionnés (riz Africains, sauvages et cultivés, riz Asiatiques d'intérêt pour nos partenaires et projets) sont en cours de séquençage profond (*Illumina* 100 PE, 25x) par le Génoscope. A cette occasion, nous allons séquencer la majeure partie de la collection du programme MENERGEP, une partie des variétés NERICA ainsi que les adventices et lignées camarguaises.

— **Séquençage des NAMs** : en collaboration avec le CIAT et AfricaRice, l'IRD a développé une série de lignées (4000) de type *Nested-Association Mapping* (Figure 2.1). Ces 4000 individus issus de croisements inter-sous-spécifique *indica* x *japonica* sont actuellement séquencés à Yale par le laboratoire de Steve Dellaporta sur la technologie *RAD-Seq*, et le seront en 2015 en *Whole-Genome Sequencing* 1x par le programme IRIGIN.

— **Séquençage pour la Sélection Génomique** : le programme conjoint entre le CIAT et le CIRAD sur la mise au point et la validation de la sélection génomique demande une validation par séquençage de différents individus, avec 200 individus en 3x et 700 en 1x (Figure 2.2).

— **Séquençage de population d'intérêt** : différents programmes de l'IRD et du CIRAD ont besoin d'information de séquençage bas-débit (1x) sur un grand nombre de lignées (650) pour obtenir des informations de cartographie génétique ou de GWAs.

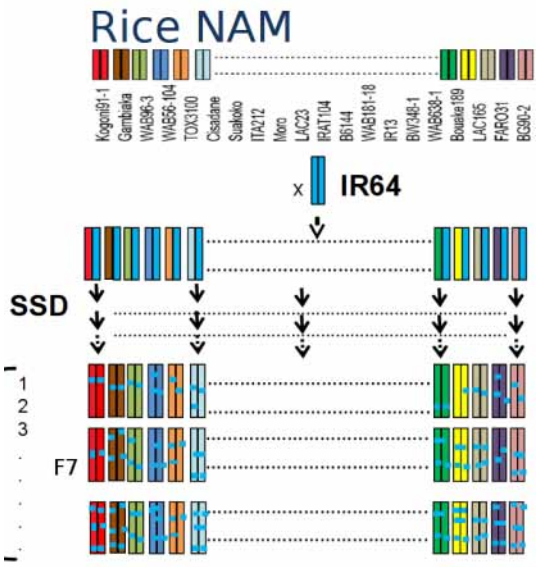

FIGURE 2.1 – Schéma de croisement des NAMs de riz

.

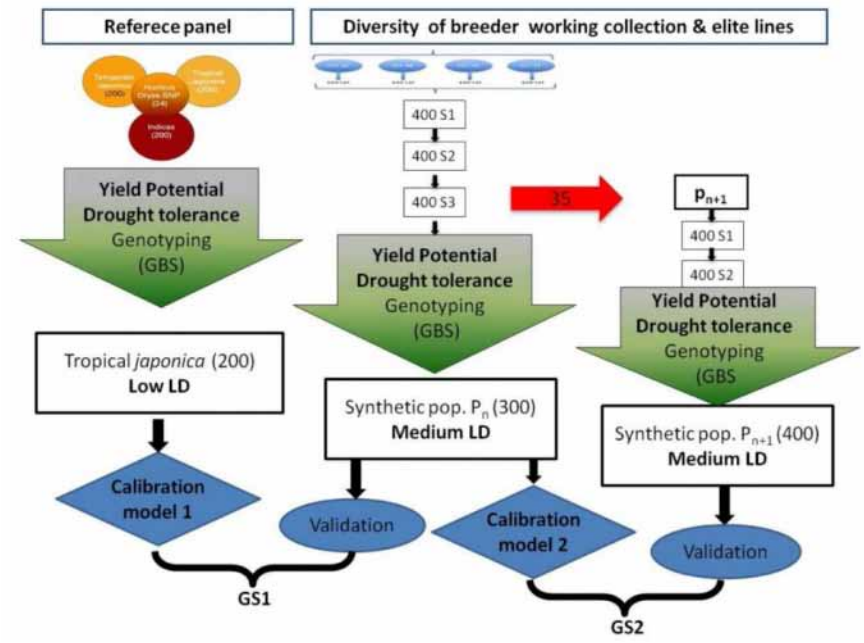

FIGURE 2.2 – Schéma de la sélection génomique. De N. Ahmadi.

.

2.2 Pan-Génome, *Core*-Génome et Eléments Transposables

2.2.1 Le concept d'espèce à l'ère de la génomique massive

D'après Mayr [MAYR, 1963], les espèces eucaryotes sont des groupes de population naturelles, effectivement ou potentiellement interfécondes, qui sont génétiquement isolées d'autres groupes similaires. En général, on considère une espèce biologique aujourd'hui comme une communauté reproductive de populations. Néanmoins cette définition, si elle fonctionne assez bien chez les animaux, est beaucoup plus difficile chez les plantes, où se produisent très souvent des hybridations. Dans cet ordre d'idée, on joint parfois à cette définition des notions d'écologie et de géolocalisation des groupes.

Toutes ces définitions considèrent généralement que le contenu génomique reste identique entre les individus de la même espèce. Dans le monde procaryote, cette vue est largement battue en brêche particulièrement dans le domaine des bactéries [Tettelin et al., 2008], avec l'identification des ***Core*** et **Pan**-génomes (Figure 2.3).

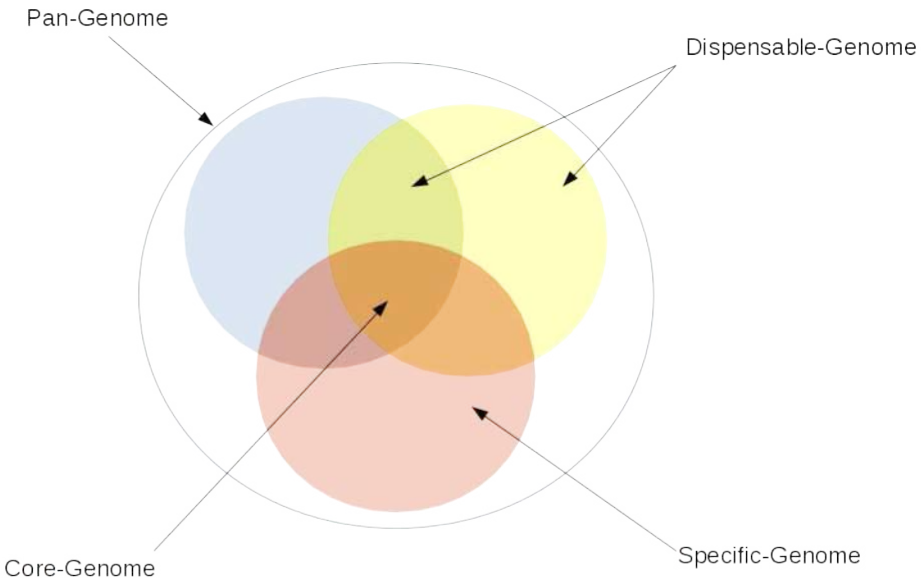

FIGURE 2.3 – Représentation schématique des Pan, *Dispensable*, *Specific* et *Core*-génomes.

.

Des individus différents (ici 3, bleu, jaune et orange, Figure 2.3) de la même espèce partage un nombre donnée de gènes/séquences génomiques (le ***Core*-génome**), qui représente la séquence *cœur* du génome de l'espèce, sa carte d'identité. Chaque individu possède une partie **Spécifique** de son génome, qui lui est unique, et partage le reste avec certains autres individus, le ***Dispensable*-génome**. L'assemblage de tous les *Dispensables* & *Specific* et du *Core*-génome compose le **Pan-génome**, *i.e.* le contenu réel en terme de séquence du génome de l'espèce entière.

2.2.2 Le Pan-Génome chez les *Poaceae*

Dans le monde bactérien cette notion très répandue est utilisée entre autre dans l'industrie, et permet aussi d'expliquer les différences de virulence de certaines souches pathogènes [Tettelin et al., 2008]. Au niveau eucaryote, quelques études utilisent les termes de Pan-génomiques, parfois à tort (chez l'homme souvent [Li et al., 2010], où ce terme est confondu avec la variabilité génétique), souvent à

raison, comme chez le *Bombyx* [Xia et al., 2009], le maïs [Hirsch et al., 2014, Morgante et al., 2007], le riz [Sakai et al., 2014, Schatz et al., 2014]...

Dans ces plantes supérieures comme chez les bactéries, plusieurs analyses ont montré des liens entre les variations pan-génomiques et les phénotypes, phénotypes non-liés à des changements épigénétiques ou alléliques. Des données récentes laissent supposer que dans les *Poaceae*, ces variations pan-génomiques seraient fortement liées à des contraintes adaptatives locales (comme dans le cas du gène *Pup1* du riz Asiatique *Oryza sativa*, [Schatz et al., 2014]), et aux phénomènes de vigueur hybride. La majeure partie des données sur les Pan-génomes de *Poaceae* sont disponibles chez le maïs, où cette notion a été utilisée pour expliquer en partie la vigueur hybride. Ainsi, dès 2007, Morgante propose une vision Pan-génomique du maïs [Morgante et al., 2007], tout du moins au niveau des éléments transposables, où seulement 75% du génome serait partagé entre deux variétés. Le même type de résultat est obtenue en 2014 par Hirsch et al.[Hirsch et al., 2014], au niveau du transcriptome de 503 lignées de maïs, avec seulement 16.3% du transcriptome commun à l'intégralité des lignées. Au niveau du riz Asiatique, Schatz et al. identifient dans le génome complet de 3 variétés [Schatz et al., 2014], et près de 300 gène spécifiques d'une variété par rapport à une autre (Figure 2.4).

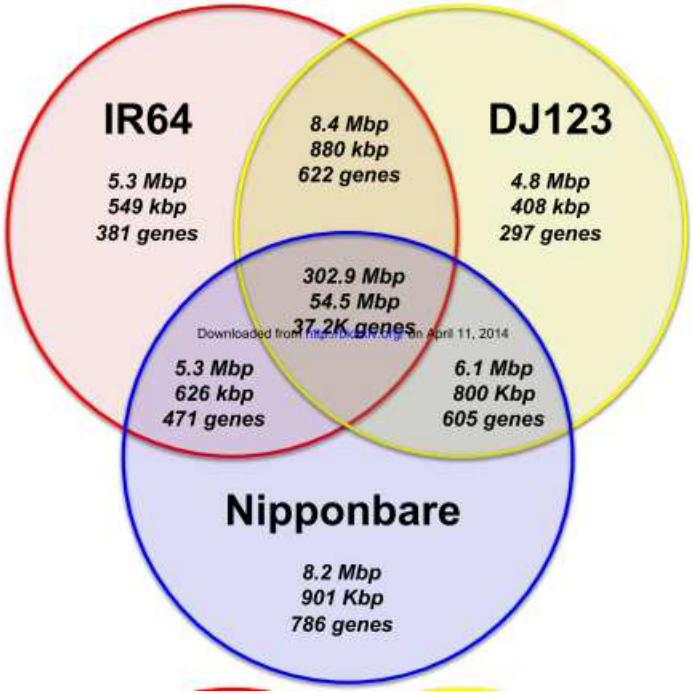

FIGURE 2.4 – Représentation des *Core*, *Dispensable* et *Specific* génomes, pour 3 variétés de riz Asiatique. La première valeur représente le nombre de bases total, la deuxième le nombre de bases en zone exonique, et la troisième le nombre de gènes.De [Schatz et al., 2014].

Toujours dans le riz Asiatique, le groupe de Jaizhong Wu a montré aussi le même résultat sur une autre variété de riz [Sakai et al., 2014], toujours en génome complet, avec presque 10% du génome (40 Mb sur 400) de différence entre Nipponbare et Kasalath. Cette analyse a aussi démontré que l'identification du polymorphisme contenu dans l'espèce *O. sativa* entière ne pouvait se faire en utilisant Nipponbare comme seule référence. Au niveau des éléments transposables du genre *Oryza*, des analyses, sur *Tos17* par exemple, ont montré que le nombre de copies variait au sein des espèces aussi, avec une variation de 1 à 11 copies de *Tos17* pour le seul riz Asiatique [Petit et al., 2009].

Ces différentes analyses montrent bien que les différences entre individus d'une même espèce ne sont pas limitées à des mutations ponctuelles, mais bien à leur contenu. Se pose alors la question de définir pour une espèce donnée ce que sont ces *Core* et Pan-génomes.

2.2.3 Étude Pangénomique des riz africains dans le cadre de la domestication

Au niveau des riz Africains *O. glaberrima* et *O. barthii*, je met en place une analyse de ces composants pan-génomiques, à deux niveaux, géniques et non-géniques. La collection que nous avons constitué dans le cadre du programme MENERGEP est actuellement en cours de séquençage profond (*Illumina* 100 PE, 25x) au Génoscope par le programme IRIGIN. Dans ce cadre, nous allons obtenir les données génomiques suffisantes pour *O. glaberrima* et *O. barthii* pour tester différentes hypothèses liées au concept de Pan-génome.

Les analyses suivantes seront réalisées intra-espèce, mais surtout inter-espèce : nous avons la possibilité d'analyser l'évolution du Pan-génome au sein de cette évènement de spéciation qu'est la domestication, et nous en tirerons de très importantes informations sur la nature et le type de modifications impliquées dans cette spéciation.

Ce travail fait déjà l'objet d'une thèse que j'encadre complètement (Mlle Cécile Monat) pour l'identification du *Core*-génome, et a été proposé à l'appel d'offre ERC *Consolidator grant* (refusé), ainsi qu'à l'ANR 2014 dans l'appel JCJC (en cours ; je suis porteur dans les deux cas). L'analyse sera effectué sous ma direction, avec la collaboration de Mmes Cécile Berthouly-Salazar (analyse évolutive des gènes) et Christine Tranchant (bioinformatique et variations).

Composants géniques du Pan-génome Pour le moment, peu d'analyses exhaustives des différences en termes de *Gene Ontology* ont été réalisée entre les différents compartiments du Pan-génome. Seule l'analyse de Hirsch [Hirsch et al., 2014] sur le maïs laisse à penser que les différences au niveau des *Specific*-génomes sont dues à des adaptations locales. Les données que nous allons obtenir dans le cadre du projet IRIGIN vont nous permettre de comparer les contenus en *GO* des *Core* et *Dispensable* génomes, et donc de vérifier cette hypothèse pour les riz Africains. Pour déterminer le *Core*-génome, nous utilisons nos assemblages de références de *O. glaberrima* ainsi que celui produit par le projet OMAP. L'idée est de croiser les données issues des mapping avec les différentes annotations possibles, et ainsi de déterminer une matrice de type présence/absence dans les différentes variétés, et donc de définir le *Core*-génome comme le set minimal de gènes présents chez tous les individus de l'espèce.

La limite de détection de présence/absence ici repose sur un compte de *reads* présents sur chacune des références. Or, ce compte peut être biaisé par le taux de répétition de cette séquence, comme par le taux de divergence. De plus, la présence de quelques *reads* peut indiquer non seulement un problème de mapping mais aussi une délétion partielle de cette séquence. Nous allons donc tester plusieurs scénarios, avec différents seuils et un contrôle en *wet lab* sur une dizaine de locus pour fixer une limite. Une fois cette limite obtenue, nous obtiendrons le set minimal de gènes présent dans le *Core*-génome, ainsi qu'un certain nombre de gènes appartenant aux différents *Dispensable* et *Specific*-génomes. Une autre analyse sur cette fois les données non-mappées nous permettra d'avoir accès aux gènes potentiellement nouveaux dans d'autres génomes que les références utilisées. Des outils tels que *NovelSeq* [Hajirasouliha et al., 2010], *BreakDancer* [Chen et al., 2009], *VariationHunter* [Hormozdiari et al., 2010], *InGap-SV* [Qi and Zhao, 2011], et d'autres seront utilisés. Nous agrandirons ainsi la taille de ce Pan-génome pour *O. glaberrima* et *O. barthii*. Une fois les données obtenus sur les différents compartiments, nous commencerons par une analyse classique de type *Gene Ontology* en cherchant des différences de répartitions *GO* entre ces compartiments, en qualitatif et quantitatif : fonction métabolique générale, fonction biochimique, expression/fonction cellulaire, tissulaire,... Dans le même temps, nous ferons une analyse de variabilité génique (SNP/InDel) entre ces mêmes compartiments, pour identifier des pressions de sélection différentielles possibles.

Notre hypothèse initiale est que le *Core*-génome va présenter un biais de *GO* en faveur des gènes de type *housekeeping* et essentiels au développement, alors que les *Dispensable* et *Specific* génomes seront plus liés à des gènes d'adaptation locale, voire à de nouveaux gènes qui ne se sont pas encore répandu dans la population (par manque de pression de sélection par exemple). De même, nous nous attendons à une variabilité moindre des gènes du *Core*-génome par rapport aux autres compartiments. En interspécifique, nous attendons une réduction de la taille du Pan-génome global pendant la domestication, mais avec une conservation quasi-complète du *Core* : la réduction globale s'effectue probablement par une réduction plus importante des compartiments *Dispensable* et *Specific*. De même nous attendons une réduction de la variabilité globale des allèles, comme dejà montré dans d'autres études sur la

domestication.

Composants non-génique du Pan-génome De nombreuses analyses se servent de différences d'insertion d'éléments transposables comme marqueurs moléculaires, avec beaucoup d'approches différentes (*S-SAP, IRAP, REMAP, TD, RBIP...*) [Kalendar et al., 2011], généralement dans le but de densifier des cartes génétiques, parfois dans l'idée de faire de la phylogénie moléculaire. Toutes ces différences d'insertions sont partie intégrante du Pan-génome, et représentent une sous-partie de la fraction *Dispensable* des TEs. Mais il est aussi possible d'aborder le Pan-génome TE non pas sur le simple aspect des insertions mais aussi sur la présence/absence des familles elles-mêmes. En premier lieu, nous allons dresser la carte des insertions communes (*Core*-génome insertionnel), puis des familles présentes chez tous les individus analysés (*Core*-génome présentiel). L'analyse insertionnelle sera réalisé en combinant les informations de mapping sur nos références de *O. glaberrima* et celle de OMAP avec les annotations, comme pour les gènes. Ensuite, nous identifierons les familles présentes simplement en scorant celles qui ont au moins une insertion.

L'étude du *Dispensable* insertionnel et présentiel sera effectué *via* les anomalies de mapping sur les références, en utilisant encore les données issues des outils de type *BreakDancer*, *VariationHunter* ou *InGap-SV*, ainsi que des outils "*home-made*", comme dans [Sabot et al., 2011]. L'idée est de pouvoir identifier à la fois le locus de variation (anomalie de mapping avec un seul *mate* mappé en bordure, mapping partielle d'un certains nombre de *reads*,...), mais aussi l'élément causal (et une reconstruction de la séquence si possible). Pour les éléments connus, nous utiliserons les bases de données TE riz, pour les autres, nous utiliserons la méthodologie que nous avons décrit dans [Wicker et al., 2007]. Les éléments du *Core*-génome seront probablement des insertions anciennes d'éléments communs au riz Africain sauvage et à son descendant cultivé, alors que les copies des *Dispensable* et *Specific* génomes seront des insertions récentes voire de nouveaux éléments.

Sortie de cette "simple" description de variation d'insertion ou de présence, nous allons pouvoir étudier les effets de la domestication/spéciation sur les familles d'éléments transposables, ainsi qu'étudier les relations possibles entre les différents groupes. Ainsi il est possible que certaines insertions voire familles d'éléments soient présentes dans le *Dispensable* dans le génome de l'espèce sauvage, mais dans le *Core* de l'espèce cultivé. Si un tel cas était identifié, nous aurions identifié le pool de sauvages à l'origine des cultivés pour le riz Africain.

Enfin, une partie des données servira aussi pour analyser l'évolution des élément non-autonomes au cours d'une spéciation. Nous étudierons leur taux d'apparition, la fréquence de transposition, l'efficacité apparente de leur parasitisme sur les éléments actifs, et enfin leur impact au global sur les génomes au cours de la spéciation.

Pour aller encore plus loin, nous pourrons analyser l'évolution des TEs sur une plus longue durée dans le groupe des *Sativa*, en utilisant les données disponibles chez le riz Asiatique (3000 génomes de riz,...). Nous aurons ainsi accès à une énorme source d'informations tant en termes de données brutes que de diversité génétique. Diverses hypothèses sur l'évolution des TEs au cours de la spéciation pourront être testés : effet fondateur, apparition de nouvelles copies-maîtres, génération de nouvelles familles, etc...

2.2.4 Attendus & Retombées

L'étude du Pan-génome de *O. glaberrima* et de son ancêtre *O. barthii* nous permettra de mieux comprendre les effets de la spéciation, et plus particulièrement de la domestication, au niveau génomique globale, et ce à l'échelle d'une espèce entière, ayant peu de diversité intrinsèque. De données immédiatemment utilisables en amélioration variétale à une meilleure compréhension de l'évolution des éléments transposables, cette analyse nous fournira de très nombreux résultats dans beaucoup de domaines :

— SNP et marqueurs moléculaires de type RBIP utilisables en amélioration variétale et cartographie génétique ;

— Marqueurs de sélection pour identifier les zones du génome soumises à sélection lors de la domestication et identification des allèles de domestication ;

— Nouvelles allèles pour des gènes d'intérêt ;
— Nouveaux gènes d'intérêt ;
— Identification de variants structuraux et mise en relation avec des phénotypes d'intérêt ;
— Identification de nouvelles familles d'éléments ;
— Meilleure compréhension des interaction autonomes *vs* non-autonomes ;
— Mise en évidence des différences et du rôle des différents compartiments du Pan-génome.

2.3 Origine des Riz Adventices et Syndrome de Dé-domestication

2.3.1 Les riz adventices

Dans tout champs, il pousse des plantes non désirées, souvent appelés adventices. Ce sont généralement des plantes envahissantes différentes de l'espèce cultivée (comme le coquelicot ou bien l'amarante ou le chardon), ou bien des individus hybrides issues de croisement de l'espèce cultivée avec des sauvages environnants. Parfois l'origine de ces adventices est encore plus complexe, et semble liée à des instabilités génomiques.

Les adventices sont généralement éliminées par l'action de produits phytosanitaires spécifiques et par une élimination manuelle. Ces deux méthodes sont assez efficaces, mais les produits phytosanitaires de ce type sont chers, très polluants et nocifs pour l'environnement, et sont en général contournés rapidement par les souches résistantes d'adventices. Dans le cas des traitements manuels, c'est une action de longue haleine, très souvent complexes et coûteuses. Dans les pays du Sud, les adventices posent ainsi de réels problèmes aux agriculteurs locaux ; en Europe, l'apparition de résistance est le problème majeur. Pour le riz, Asiatique comme Africain, il existe une très forte pression d'adventices, avec une contamination entre 40 et 85% des champs par des riz dits adventices (Figure 2.5).

FIGURE 2.5 – Riz adventice dans un champs de riz Asiatique de variété élite.

Ces riz adventices (aussi appelé *Crodo* riz rouge ou *O. sativa f. spontanea*) se caractérisent par une plus forte vigueur à la croissance, une plus grande taille, de plus petits grains, un péricarpe souvent rouge (Figure 2.6), un égrenage fort et une dormance longue. Ce retour à des traits sauvage est appelé aussi parfois "syndrome de dé-domestication". De plus, ces riz adventices sont fertiles avec le riz cultivé. Ainsi, ils poussent plus vite que les riz cultivés, plus haut, et sont très égrenants. Les grains tombés au sol peuvent germer l'année suivante, envahissant le champ petit à petit. Les agriculteurs éliminent généralement ces adventices (et les autres) *via* une action manuelle, et limitent leur apparition en appliquant une strict tri des stock de graines. Le péricarpe des riz adventices étant souvent rouge, les semences de riz sont passées sur un banc chromatique, qui ne conserve que les lots ayant un taux de grains rouges inférieur à 1 pour 500 grains blancs. De même, tout un système de pratiques culturales est mis en place pour éviter les contaminations inter-champs ou annuelles : nettoyage des engins,

changement de bottes et chaussures, filtrage des eaux d'irrigation, assolement, rotation des cultures... Malgré tout, la très grande capacité de dormance des riz adventices (jusqu'à 7 ans) leur permet de repousser et réinvestir un champs en l'espace d'une seule année. Ainsi, les pertes économiques induites par les riz adventices sont énormes, juste après la sécheresse et avant les pathogènes au champs ! Leur impact est mondial, et chaque année de nouveaux riz adventices apparaissent au champs.

FIGURE 2.6 – A gauche des grains de riz normaux, à droite des grains de riz adventices. Du *Lawton-Rauh Laboratory*.

Dans plusieurs cas, leur origine a pu être identifié dans des stocks de graines mal triées, et issues de croisement entre le riz cultivé et des espèces sauvages, ou entre les différentes sous-espèces de riz Asiatique, *japonica* et *indica* [LONDO and SCHAAL, 2007, Song et al., 2014]. En effet, les croisements interspécifiques (ou inter-sous-spécifiques) non désirés ont produit en première génération des plantes anormales, qui ont pu se multiplier rapidement et envahir les champs, notamment aux USA.

Néanmoins, dans certains cas la situation est moins claire. Ainsi, les analyses de Cao et al [CAO et al., 2006] ou de Qiu et al [Qiu et al., 2014] ont montré que certains riz adventices chinois ne sont pas issus de croisements avec les riz sauvages apparentés, mais sont en fait issus directement des variétés cultivées elles-mêmes. De même, les groupes américains de Olsen [Gross et al., 2010] et Caicedo [Thurber et al., 2010] ont découverts que les caractères responsables de ce syndrome de dé-domestication, et en particulier l'égrenage et la couleur du péricarpe, n'étaient pas liés aux gènes sélectionnés contre ces phénotypes dans le cadre de la domestication. En effet, les gènes identifiés ne sont pas les gènes connus chez les sauvages comme étant responsables de ces phénotypes : l'origine de ces adventices est donc clairement différente de celle de ceux issus de croisements avec les sauvages.

Le cas camarguais En Camargue, autour de Arles, se situe la seule zone rizicole de France. La production y est modeste, principalement des riz Asiatiques de type *japonica*, dans de petites exploitations. Cette culture est certes mineure, mais reste emblématique de la zone camarguaise : elle permet l'entretien des canaux, le recul des marais salants et des zones infestantes en moustique, ainsi que la désalinisation des sols. De plus, la filière elle-même emploie plus de 1000 personnes sur le bassin de Arles. Le Centre Français du RIZ (CFR), une association loi 1901, produit de nouvelles variétés adaptées au marché semencier français, entre autres activités. Le CFR gère aussi les risques d'enherbement par les adventices (Figure 2.7), et proposent des approches pour éviter ces contaminations.

Leur système de sélection et de purification des lignées est de très haute qualité, avec des contraintes et des normes de haut niveau. Les grains en nurserie sont plantés manuellement sur un rouleau qui est ensuite déroulé dans la rizière de la nurserie. Chaque grain est étiqueté, et tout ce qui pousse sur les parcelles d'essais a un pedigree bien connu. Néanmoins, malgré tous les soins apportés à la pureté des échantillons et à la sélection, il apparait régulièrement des adventices de type riz, même en nurserie (Figure 2.8).

On peut d'ors et déjà ici éliminer le risque de contamination par des riz sauvages (inexistant en France). Les stocks de graines du CFR sont validés sur banc chromatique mais aussi manuellement lors de l'installation sur le rouleau, et l'on peut donc aussi éliminer le risque de contamination par

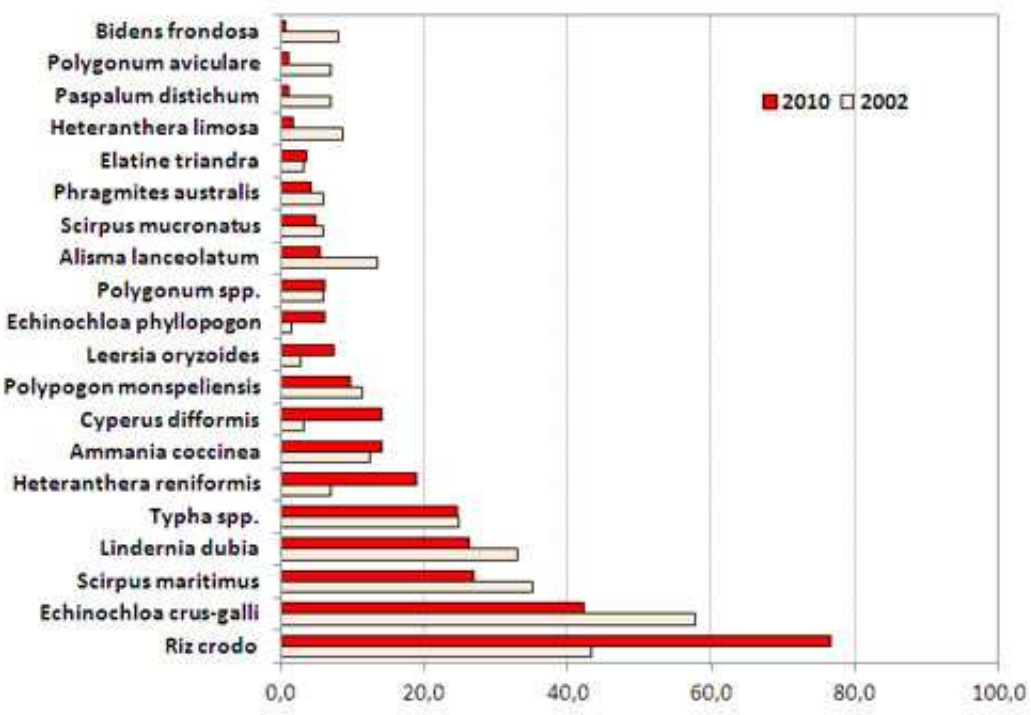

FIGURE 2.7 – Variation des taux d'enherbement sur les parcelles camarguaises entre 2002 et 2010. On voit une forte augmentation des parcelles touchées par les riz adventices, de type Crodo. Du CFR.

FIGURE 2.8 – Adventice au CFR, dans une lignée F3 suivie en nurserie.

un adventice de l'année précédente. Reste la possibilité que ces riz adventices apparaissent plus ou moins spontanément dans certains fonds génétiques. Depuis quelques années, les agronomes du CFR

ont introduit dans leurs schémas d'amélioration des variétés arômatiques ainsi que des variétés *indica*. De tels croisements ont donnés lieu à des hybrides, purifiés ensuite en lignées par autofécondations successives, lignées ensuite inscrites au catalogue puis elles-mêmes utilisées en sélection. Des apports d'alléles sont aussi fréquemment effectués en ajoutant aux schémas de sélection des plantes issues des sélections italiennes ou espagnoles.

Toutes ces plantes sont donc issus de croisements inter-sous-spécifiques, et notre hypothèse est que ces croisements même provoquent des instabilités génomiques au cours des générations, jusqu'à obtenir des riz adventices en bonne et due forme. Du matériel de type adventices (Crodo) et en cours d'adventisation (crodoïforme) a été collecté par le CFR sur les 10 dernières années, ainsi que les plantes des années *n-1* et *n* correspondantes. Nous allons utiliser ce matériel pour tester nos hypothèses.

Le cas des NERICA En Afrique Sub-Sahélienne, les principales variétés de riz à haut rendement cultivées sont des variétés NERICA (*NEw RIce for Africa*). Ces variétés ont été créé au début des années 2000 par le Dr Monty Jones, agronome à AfricaRice, à partir de croisements interspécifiques entre *O. sativa* et *O. glaberrima* [Gridley et al., 2002]. Ce croisement donne des F1 stériles à plus de 99% en ovulaire, et complètement en pollinique, entre autre par l'action du gène de stérilité S^1, situé sur le haut du bras court du chromosome 6. Il est possible, en forçant des *backcross* (BC) sur un des deux parents (donneur mâle), d'obtenir après 3 de ces BC une plante présentant un niveau de fertilité pollinique et ovulaire satisfaisant pour des autofécondations (Figure 2.9). Le même résultat peut être obtenu aussi en effectuant une culture d'anthères de l'hybride F1 ou des différents BC, mais avec plus ou moins de succès.

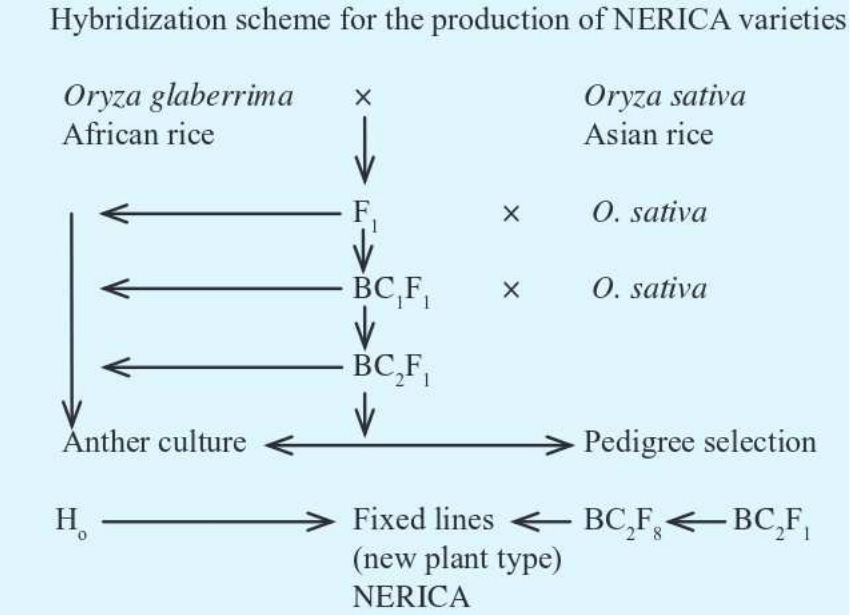

FIGURE 2.9 – Schéma de croisements, rétrocroisements et autofécondations pour la création des NERICA. De AfricaRice.

Deux générations de NERICA ont été créé : les 18 premières lignées à partir de CG14 comme donneur Africain et des lignées apparentées à WAB56104 comme donneur Asiatique, et sont considérées comme des variétés dites *upland*, ou pluviales (NERICA-1 à NERICA-18), et les 60 suivantes par un croisement entre TOG5681 comme donneur Africain et IR64 comme donneur Asiatique ; ces dernières lignées sont dites *lowland*, ou irriguées (NERICA-L-1 à NERICA-L-60). Ces différentes variétés NERICA combinent le meilleur des deux riz cultivés, à savoir une productivité accrue par rapport à *O. glaberrima* et une meilleure résistance aux conditions Africaines, en particulier les pathogènes et la sécheresse par rapport à leur parent Asiatique *O. sativa*. Ainsi, ces variétés ont été très rapidement adoptées par les paysans Africains, telles quelles ou dans les programmes de sélection nationaux.

Au Bénin, l'INRAB (Institut National de Recherche Agricole du Bénin) produit des semences de

NERICA pour le pays entier, en particulier des NERICA 1, 2 et 4. L'INRAB a mis en place ces dix dernières années un important protocole agronomique pour pouvoir suivre et améliorer la qualité de ses semences, en particulier pour réduire le taux de mélange dans les graines donnant lieu à des infestations d'adventices, associé à une préparation du sol avec de faux semis et une utilisation massive du *round-up*. Les champs semenciers sont continuellement surveillés par des paysans, qui retirent toutes les plantes suspectes manuellement. Les stocks de semences de pré-base et base sont reconstitués en suivant un schéma classique de régénération avec conservation phénotypique, puis repiquage en ligne des panicules et élimination des lignes non conforme. Les graines sont plantées en ligne, pour identifier toute intrusion de riz hors du stock initial (Figure 2.10).

FIGURE 2.10 – Champs semencier de l'INRAB pour la variété NERICA-4 en 2012, à proximité de Bohicon, au Bénin.

Malgré toutes ces précautions, en 2011, une des plus grandes parcelles semencières pour NERICA-4 a du être déclassée à cause d'une forte apparition de riz à phénotype *glaberrima* : grains noirs, glabres, plus égrenants, moins productifs. Ces adventices étaient réparties dans le champs sous forme de taches séparées, en ligne, et sans rapport avec la méthodologie d'irrigation. En 2012, le stock de graines utilisées pour régénérer les semences a eu la même origine, sans toutefois obtenir ce même type de contamination (Figure 2.10). Seuls quelques plantes anormales dans les lignées issues de la même panicules apparaissaient, avec un taux de contamination du champs final inférieur à une plante pour 1000. Aucune explication n'a pour le moment été trouvé sur l'apparition soudaine de ces adventices. Au niveau des champs de production des fermiers béninois, certaines plantes adventices apparaissaient en 2012, probablement liées à des croisements interspécifiques l'année précédentes (entre *O. sativa* et *O. barthii*, voire entre variétés *indica* et *japonica* ; Figure 2.11).

En 2013, le nord Bénin a subi une infestation massive d'adventices dans les champs de production

FIGURE 2.11 – Adventice interspécifique dans un champs de production proche de Dassa, au Bénin.

de NERICA-4, la variété la plus utilisée dans ce pays. Les pertes économiques des fermiers locaux ont été très importantes, et ces agriculteurs en reportent la responsabilité sur l'INRAB, en accusant le stock de graines de contamination. Or, les graines provenaient des champs de 2012 que nous avions inspecté nous-mêmes. Les paysans du nord Bénin ne veulent plus de NERICA-4 actuellement, et plusieurs autres systèmes agronomiques nationaux ont aussi rapporté depuis 2012 des problèmes d'instabilité des riz NERICA (Togo, Ghana, Burkina Faso, entre autres).

Dans ce cadre, en assocciation avec l'INRAB (Mlle Iliyath Bello), AfricaRice (Dr Moussa Sié) et l'université FAST de Dassa/Abomey-Calavi (Dr Gustave Djedatin), nous avons décidé de mettre en place une étude de ces adventices, sur la même méthodologie que pour les riz camarguais. Nous avons obtenu de l'AfricaRice les graines d'origines (semences G0) des différentes variétés NERICA, ainsi que certains des adventices des NERICA *via* l'INRAB.

2.3.2 Méthodes envisagées

Dans ces deux cas camarguais comme béninois, nous tablons sur une instabilité de type génomique, avec une activation ou réactivation des éléments transposables. pour ce faire, nous allons séquencer les individus adventices et les comparer à des plantes-soeurs n'ayant pas subi ces variations.

La première chose sera d'identifier de vrais adventices issus des lignées cultivées, et non pas des hybrides F1. Pour cela, nous utiliserons les marqueurs classiques à notre disposition, microsatellites et SNP, entre autres. Puis nous séquencerons dans le programme IRIGIN les individus d'intérêt et leur plante-soeur, pour scorer les différences et identifier les évènements génomiques conduisant à ce syndrome de dé-domestication à travers des analyses de mapping sur les différentes ressources génomiques de riz Asiatiques et Africains à notre disposition. Actuellement, nous avons lancé le séquençage des variants et normaux camarguais, ainsi que des principales variétés NERICA. L'identification de ces évènements conduira ensuite à une analyse basée sur l'épigénétique et le contrôle des éléments transposables dans le cadre des croisements interspécifiques et inter-sous-spécifiques, en collaboration

avec Mlle Marie Mirouze. Ce projet a déjà été déposé à l'Appel d'Offre ANR *BioAdapt* 2013 (refusé pour mauvais choix d'AO), ANR *Grand Défi Sécurité Alimentaire* 2014 (refusé pour raison inconnue) et Agropolis *OpenScience* 2013 (refusé car trop cher).

2.3.3 Attendus & Retombées pour le Sud et le CFR

Les attendus principaux au niveau scientifique seront l'identification des éléments responsables de ces évènements, que ce soient des éléments transposables (ce qui est fortement probable) ou non. Au niveau agronomique, nous pourrons ensuite dériver des données génomiques pures des marqueurs de certification des lignées. De tels marqueurs PCR simples (de type RBIP ou présence/absence) pourront être utilisés dans n'importe quel laboratoire pourvu d'une simple machine PCR et d'un appareil à gel d'agarose, de manière à pouvoir valider l'appartenance d'un stock de graines à telle ou telle variété, mais aussi à certifier sa conformité génomique.

2.4 Capture de séquences spécifiques à haut débit

2.4.1 Génotypage massive et sélection variétale au Sud

Actuellement, il est possible de génotyper en masse des individus pour une somme avoisinant les 40 euros par échantillons. Ce genre d'approche de type *Genotyping-By-Sequencing* (ou GBS) [Davey et al., 2011] permet d'avoir *via* les technologies NGS accès à plusieurs centaines de milliers de points de type SNP. Ce genre de données est très intéressante en analyses de type GWAs (*Genome Wide Association*) pour identifier des relations phénotype/génotype, ou bien en analyse d'évolution génomique globale. Néanmoins, dans le cas de l'amélioration variétale, ce genre de données ne satisfait pas forcémenent le sélectionneur. En effet, ce dernier veut avant tout connaître les allèles des plantes qu'il travaille (ou qu'il veut travailler) pour une petit nombre de gènes d'intérêt, liés à un attendu agronomique donné, et une idée plus général mais moins détaillée du fonds génétique, particulièrement dans les étapes de sélection des lignées. Or, pour le moment, la seule possibilité pour les sélectionneurs du Sud pour cela reste le séquençage à l'ancienne de type *Sanger* pour quelques gènes, et un suivi des lignées par marqueurs microsatellites. Tout ceci reste coûteux et à relativement bas débit. Dans ce cadre, en association avec l'AfricaRice (Dr Marie-Nöelle Ndjiondjop & Dr Mounirou Sow), nous avons entrepris cette année de mettre au point un système de capture spécifique pour séquencer à haut débit 200 gène sélectionnés par individu, avec un coût de rendement inférieur à 40 euros par échantillon. Nous allons développer une approche similaire à celle utilisée dans l'analyse d'exome chez l'homme, par exemple, *via* des captures de séquences spécifiques. Pour la mise au point de la technologie, son transfert à AfricaRice et son application à grande échelle, j'ai fais une demande de bourse de thèse au programme GRiSS (GRiSP Scholarship), qui a été acceptée. De plus, j'ai aussi déposé une demande de financement pour ce projet et une demi-bourse de thèse sur ce programme à l'appel d'offre *OpenScience* recherche 2014 de la Fondation Agropolis (projet #hASHTAG-RICE).

2.4.2 Méthodologies

Suite à notre expérience dans le projet Chlorodiv sur la capture de séquences, nous avons décidé de choisir d'utiliser la plateforme de capture à façon *MYbaits*, en utilisant des sondes ARN de 80mers et une couverture en 2x (Figure 2.12). Le séquençage se fera en sous-traitance sur une machine *Illumina* de type *MiSeq*, avec 100 individus à la fois.

Nous allons sélectionner 200 gènes, dans lesquels il y aura un grand nombre de gènes d'intérêt agronomique dont la séquence est disponible (ou pourra être obtenue par le programme IRIGIN). Une fois les séquences sélectionnées, nous mettrons la méthodologie au point à Montpellier sur un set de *O. glaberrima* (séquencés en haute profondeur) comme contrôle. La méthodologie validée sera ensuite transférée à AfricaRice, puis appliquée plus largement sur 2000 individus de la collection des Ressources Génétiques d'AfricaRice.

Un des points forts du projet, en dehors du coût de revient minimal (40 euros pour 200 gènes par individu), sera la mise en place d'un pipeline dédié qui permettra d'obtenir un fichier multifasta par individu pour chacun des gènes, mais aussi un tableau de type Excel/Calc donnant l'allèle pour chaque individu à chaque gène (Figure 2.13). Cela sera fait en identifiant directement à partir des variants les allèles spécifiques.

La méthode permettra aux sélectionneurs du Sud d'avoir accès en un temps record et pour un coût minime à la variabilité de leurs plantes favorites et de leurs lignées en cours de fixation, avec un résultat utilisable directement, sans avoir besoin de demander encore un travail d'analyse SNP post-séquençage.

2.4.3 Attendues & Retombées pour le Sud

Ce projet permettra non seulement de mettre au point une nouvelle technologie utile pour nos partenaires du Sud, mais aussi de fournir énormément de données rapidement sur un grand nombre d'individus de riz (Asiatique ou Africain, sauvage ou cultivé) pour des analyses évolutives et des aspects d'amélioration. Enfin, ce projet servira de plateforme de formation en bioinformatique (en

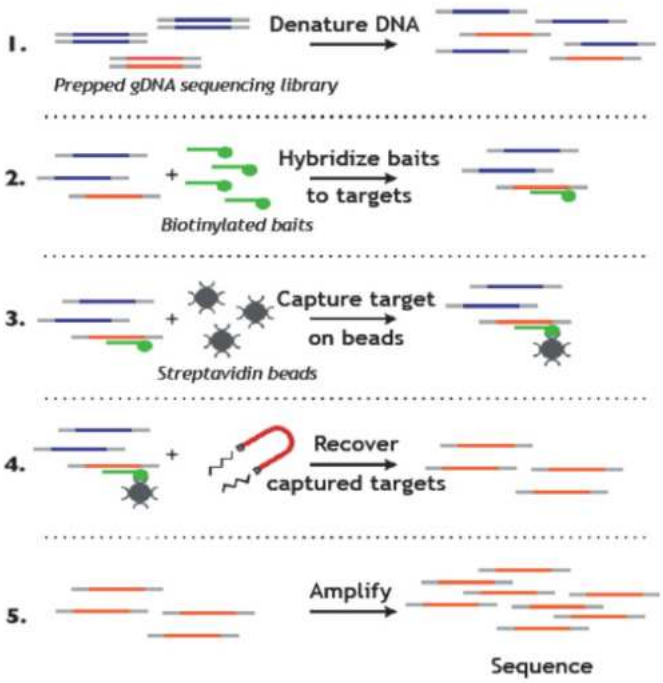

FIGURE 2.12 – Méthodologie de capture de gènes à partir de sondes *MYbaits* de type ARN. Une fois l'hybridation sonde/cible réalisée sur la banque *Illumina* complète, les fragments ciblés sont récupérés et séquencés. De *MYbaits*

Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

Liberation Sans 10

G11

	A	B	C	D	E	F	G
1	**Gene_ID**	**Allele_Sample1**	**Allele_Sample2**	**Allele_Sample3**	**Allele_Sample4**	**...**	**Allele_SampleN**
2	Gene1	var1	var1	var3	var2	...	var1
3	Gene2	a	b	b	a	...	b
4	Gene3	High	High	High	Low	...	Low
5	...	...	...	...	...	...	...
6	GeneN	1	2	2	2	...	1
7							

FIGURE 2.13 – Prototype du tableau de sortie des allèles dans les différents individus testés.

bioanalyse comme en programmation), pour le ou la Doctorant(e) du projet et aussi pour le Dr Sow d'AfricaRice.

Chapitre 3

Curriculum Vitae

3.1 Etat Civil

François Sabot

Né le 29 Août 1977, à Aurillac (Cantal, France).

Nationalité Française

Marié, 2 Enfants

3.2 Situation Actuelle

Chargé de Recherches de Première Classe (depuis 2011)

Institut de la Recherche pour le Développement IRD

UMR 232 DIADE DIversité, Adaptation, DEveloppement IRD/UM2

Equipe Génome & Développement des Riz GDR (2010-2014)/RIcE (2015-)

3.3 Formation Initiale

2004, Thèse de Doctorat : Spécialité Biologie Moléculaire Végétale, Université Blaise Pascal Clermont II. Mention Très Honorable avec Félicitations du Jury.

2001, DEA en Physiologie et Génétique Moléculaire : Université Blaise Pascal Clermont II. Mention Bien.

2000, Maîtrise de Biologie Cellulaire & Physiologie : Option Physiologie Végétale, Université Blaise Pascal Clermont II. Mention Assez Bien.

1999, Licence de Biologie Cellulaire & Physiologie : Option Physiologie Animale & Végétale, Université Blaise Pascal Clermont II.

1998, DEUG Sciences de la Vie & de la Terre : Université Jean Monnet Saint-Etienne.

3.4 Expérience Scientifique

2001-2004 : Thèse de Doctorat à l'INRA de Clermont-Ferrand, UMR Amélioration & Santé des Plantes, au sein de l'équipe Génome, sous la direction de Michel Bernard. *Caractérisation & Evolution des Eléments Transposables dans le complexe des espèces Aegilops/Triticum.*

2005-2007 : Contrat Post-doctoral au Plant Genomics Laboratory, Institute of Biotechnology, de l'Université d'Helsinki (Finlande). *Analyse et études des Eléments Transposables non-autonomes dans les génomes des Triticeae.*

2007-2010 : Chargé de Recherches IRD au Laboratoire Génomes & Développement des Plantes IRD/CNRS/UPVD de l'Université Via Domitia de Perpignan. *Evolution des Eléments Transposables du riz Oryza sativa.*

Depuis 2010 : Chargé de Recherches IRD dans l'UMR DIADE de l'IRD de Montpellier, équipe GDR. *Evolution & Structure des Génomes des riz Africains.*

3.5 Expérience en Enseignement

3.5.1 Au Nord

2001-2004 : Monitorat à l'Université Blaise Pascal ClermontII - UFR Sciences, Département Biologie Végétale. 64h/an équivalent TD

2008-2010 : Master DYNEV, module *Analyse du Génome*, Université de Perpignan Via Domitia. 30h/an équivalent TD

2010-2011 : Master APIMET, module *Génomique Végétale*, Montpellier SupAgro. 3h/an, CM.

2010-2014 : Master APIMET/SEPMET, module *BioInformatique Appliquée*, Montpellier SupAgro. 14h/an équivalent TD.

2010-2013 : Formation IRD *Introduction à la Bioinformatique*. 25h/an

3.5.2 Au Sud

2010-2014 : Master BioPharma, Université Sciences et Technologie de Hanoi, Module BP32 *Bioinformatics for Plant Genomics*. 40h/an

2011-2014 : Université Cheik Anta Diop, Dakar, Sénégal, module *iPlant*. 3h/an, CM

2013 : Faculté des Sciences et Techniques de Dassa/Abomey-Calavi, Bénin, *Ecole chercheur en Bio-Informatique*. 45h.

2014 : Ecole thématique *SudBiotech* CIRAD/IRD/Université d'Abomey-Calavi, Bénin. 40h.

Chapitre 4

Activités d'Encadrement, de Gestion de Projets et Administratives

4.1 Post-Doctorants et Chercheurs Accueillis

2014-2017 : Dr Gustave Djédatin, accueilli dans le cadre d'une BEST (Bourse d'Échange en Sciences et Technologies), sur l'étude des structures génomiques des NERICA. 3 mois/an.

4.2 Doctorants

2005-2007 : Co-encadrement de deux Doctorants (Marko Jäskelaïnen et Jääko Tanskanen), *Plant Genomics Laboratory*, Université d'Helsinki, Finlande, sur l'étude des éléments transposables non-autonomes de la famille des *BARE*.

2013-2016 : Encadrement complet d'une Doctorante, Mlle Cécile Monat, UMR DIADE, IRD Montpellier, sur l'étude des *Core-* et Pan-génomes des riz Africains.

2015-2018 : Co-Encadrement d'un ou d'une Doctorante dans le cadre du programme de capture de séquence, en collaboration avec le Dr Marie-Noëlle Ndjiondjop.

4.3 Etudiants de Master

2004 : Encadrement d'une élève ingénieur de dernière année de l'École Polytechnique de Lisbonne, sur l'activité des éléments transposables dans les blés synthétiques. 6 mois.

2009 : Encadrement d'un étudiant marocain du Master2 Dynev de l'Université de Perpignan, sur la détection des éléments transposables actifs du riz Asiatique *via* une approche par reséquençage. 6 mois.

2012 : Encadrement d'une étudiante du Master1 PCM option Bioinfo de l'Université Blaise Pascal Clermont II, sur la mise en place d'un service web d'annotation automatique des rétrotransposons à LTR. 3 mois.

2013 : Encadrement d'une étudiante du Master2 PCM option Bioinfo de l'Université Blaise Pascal Clermont II, sur l'assemblage du génome des riz Africains par une méthode de séquençage hybride *Illumina/PacificBiosciences*. 6 mois.

2014 : Encadrement d'un étudiant togolais du Master2 Bioinfo de l'Université de Rennes 1, sur l'assemblage et la détection de SNPs dans les génomes chloroplastiques. 6 mois.

4.4 Autres étudiants

2014 : Co-encadrement d'un stagiaire BTS biotechnologie de Nîmes, avec travail au laboratoire et en serre. 6 semaines.

4.5 Ingénieurs et Techniciens

2012-2014 : Encadrement du travail d'une IE en CDD, dans le cadre du projet MENERGEP sur l'analyse de la variabilité des riz Africains. 18 mois.

2010-2014 : Encadrement partiel des serristes IRD dans le cadre du projet MENERGEP, PanGenome et riz adventices.

2005-2007 : Encadrement partiel des Techniciennes de laboratoire, *Plant Genomics Laboratory*, Université d'Helsinki, Finlande.

4.6 Jurys - Comités - Consultance

4.6.1 Jurys

Recrutement de Maître de Conférence : 2

Recrutement d'Ingénieur en poste : 1

4.6.2 Comités - Consultance

Comité de Thèse : 3

Consultance pour le privé : 3

Je suis aussi régulièrement consulté par mes différents collègues de l'unité ou d'autres UMRs pour des informations relatives à des montages de projets touchant à la bioinformatique, sur des aspects scientifiques et techniques, ainsi que sur les objectifs et suivis de stages de différents niveau (L3 à M2 et thèse, hors comités). Enfin, j'ai souvent affaires à la DSI et à l'IS (informatique scientifique) sur tous les aspects institutionnels de stockage et calcul : choix et validation des méthodologies, nouvelles approches, prospective,...

4.7 Groupe d'études et Développement

2014 : Initiateur du *Dojo Code* de la plateforme collaborative SouthGreen, dédié à la mise en place de pipeline d'analyse NGS. Mise en modules, gestion collaborative (*SVN*, *Git*), mise en place de tests unitaires.

2010-2012 : Membre régulier des ArcadThons, pour la mise en place des méthodes et l'analyse des données NGS issues du programme Arcad de la Fondation Agropolis.

4.8 Responsabilités Administratives

2011-2014 : Responsable du Site Genetrop, plateforme des laboratoires des UMRs DIADE et RPB.

2010-2014 : Membre élu du Conseil d'UMR de DIADE.

4.9 Activités de *Reviewing* et éditoriales

Reviews J'ai été *reviewer* pour différents journaux internationaux tels que *PNAS*, *Genome Research*, *The Plant Journal*, *MBE*, *JME*, *BMC Genomics*, *Mobile DNA*, *Genetica*, *MGG*, *TAG*,...

Editorial Board Depuis 2013, je suis éditeur associé au journal *Genetica*, avec une activité de moyenne de 2 à 3 articles par mois à gérer.

4.10 Organisation de Colloques

2011 Colloque DIADE "Eléments Transposables des les Plantes Tropicales & Méditerranéennes", Montpellier, France

2013 15e Congrès National sur les Eléments Transposables CNET, Montpellier, France

4.11 Autres activités de Diffusion de l'information scientifique auprès du Grand Public

2010 : Journées de la Science, Perpignan, Conférence sur la taille et la structure des génomes

2014 : Utilisation des réseaux sociaux (facebook ou autres) pour le projet Framboisine

2014 : Présentation aux Mardis de l'IRD sur la bioinformatique

4.12 Dépôts et Gestion de Projets

4.12.1 Projets déposés mais non retenus

— 2005

  — Porteur : 1 (*Marie Curie Fellowship*)

- 2007
  - Porteur : 1 (*University of Helsinki*)
  - Collaborateur : 1 (*National Research System of Finland*)
- 2009
  - Collaborateur : 2 (ANR & Génoscope)
- 2011
  - Porteur : 1 (ANR)
  - Collaborateur : 3 (ANR, Grand Emprunt)
- 2012
  - Porteur : 2 (Fondation Agropolis, PPR SREC IRD)
  - Collaborateur : 2 (Fondation Agropolis, ANR)
- 2013
  - Porteur : 4 (ANR, Fondation Agropolis, JEAI IRD, PPR SREC IRD)
- 2014
  - Porteur : 2 (ANR, ERC)
  - Collaborateur : 2 (Fondation Agropolis, 7e PCRD)

4.12.2 Projets soumis 2014

- Porteur : 3 (AI IRD, ANR JCJC, Fondation Agropolis)
- Collaborateur : 1 (Fondation Agropolis)

4.12.3 Projets acceptés

- 2005
  - Porteur : *International post-doctoral position at the University of Helsinki*, 2 ans de Post-Doc.
- 2010
  - Porteur : Action Incitative IRD : mise en place d'un cluster de calcul en Bioinformatique. 10 000 euros
- 2011
  - Porteur : Action Incitative IRD : Mise en place d'un réseau de surveillance des NERICA au Bénin. 8 000 euros
  - Collaborateur
    - GRiSP : *Focal point* pour la génomique à l'IRD
    - MENERGEP : Responsable du WorkPackage 1. 400 000 euros, donc 90 000 au WP1
- 2012
  - Porteur : Programme IRIGIN, Appel d'Offre *Massive Sequencing* de *France Génomique*. ca 2 millions d'euros.
  - Collaborateur
    - IBC : Membre du WP5
    - EvoRepRice : Projet Fondation Agropolis/Fondazione Cariplo, Appel D'Offre FIRST. Responsable bioinformatique. 450 000 euros
- 2013
  - Porteur

- Projet GLASS : Projet IRD/AfricaRice sur l'assemblage des génomes de riz Africains. 65 000 euros.
- BEST sur la génomique des NERICA : Appel D'Offre IRD. Accueil de chercheur du Sud 3 mois par an pendant 4 ans.

- Collaborateur
  - AfriCrop : Projet ANR. Responsable du WorkPackage Bioinformatique. 698 000 euros.
  - Chlorodiv : Appel d'Offre *OpenScience*, Fondation Agropolis. Responsable Bioinformatique. 45 500 euros.

- 2014
  - Porteur : Projet SPIRALES Framboisine, Appel d'Offre IRD. 8 000 euros.

Chapitre 5

Production Scientifique

5.1 Articles Internationaux dans des revues à Comité de Lecture

Les papiers signalés par un symbole "†" ont été réalisés avec un ou une Doctorante ou Master que j'ai encadré ou co-encadré.

5.1.1 Publiés

25 Papiers, dont 13 en premier auteur et un en dernier, plus 2 réponses (**Nature Reviews Genetics**) et une présentation/pre-proceeding (**Nature Publishing Group**).
H-Index 13 (25 novembre 2014) - i10-Index 16 (25 novembre 2014)

Sabot F, Simon D, Bernard M : Plant transposable Elements, with and emphasis on grass species. **Euphytica** 2004, 139(3),227-247

Sabot F, Sourdille P, Bernard M : Detecting specific repeated sequences in large, complex genomes by using representative difference analysis and double-probe verification. **Plant Molecular Biology Reporter** 2004, 22(1), 91

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

Chantret N, Salse J, **Sabot F**, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier M-F, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B : Molecular basis of evolutionary events that shaped the *hardness* locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). **Plant Cell** 2005, 17 :1033-1045

Chantret N, Salse J, **Sabot F**, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier M-F, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Leroy P, Bernard M, Chalhoub B : Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. **J. Mol. Evol** 2008, 66 :138-150 - Co-premier auteur

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

Sabot F, Sourdille P, Bernard M : Advent of a new retrotransposon structure : the long form of the *Veju* elements. **Genetica** 2005, 125 :325-332

Sabot F, Sourdille P, Chantret N, Bernard M : *Morgane*, a new LTR retrotransposon group, and its subfamilies in wheats. **Genetica** 2006, 128 :439-447

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

†**Sabot F**, Kalendar R, Jääskeläinen M, Wei C, Tanskanen J, Schulman AH : Retrotransposons : metaparasites and agents of genome evolution. **Israel Journal of Ecology & Evolution** 2006, 52(3-4), 319-330.

Sabot F, Schulman AH : Template switching can create complex LTR retrotransposon insertions in *Triticeae* genomes. **BMC Genomics** 2007, 8 :247

†Tanskanen JA, **Sabot F**, Vicient C, Schulman AH : Life without GAG : the *BARE*-2 retrotransposon as a parasite's parasite. **Gene** 2007, 390 :166-174 - Co-premier auteur

Wicker T, **Sabot F**, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH : A unified classification system for eukaryotic transposable elements. **Nature Review Genetics** 2007, 8 :973-982 (plus **2 answers associated**) - Co-premier auteur

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

Petit J, Bourgeois E, Stenger W, Bès M, Droc G, Meynard D, Courtois B, Ghesquière A, **Sabot F**, Panaud O, Guiderdoni E : Diversity of the *Ty*-1 copia retrotransposon *Tos17* in rice (*Oryza sativa* L.) and the AA genome of the *Oryza* genus. **Mol. Genet. Genomics** 2009, 282 :633-652

Picault N, Chaparro C, Piegu B, Stenger W, Formey D, Llauro C, Descombin J, **Sabot F**, Lasserre E, Meynard D, Guiderdoni E, Panaud O : Identification of an active LTR retrotransposon in rice. **The Plant Journal** 2009, 58 :754-765

Roulin A, Piegu B, Fortune PM, **Sabot F**, D'Hont A, Manicacci D, Panaud O : Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in *Poaceae*. **BMC Evol. Biol** 2009, 9 :581

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, **Sabot F**, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, et al. : The genome of *Theobroma cacao*. **Nature Genetics** 2011, 43 :101-108 (plus a **Nature Publishing Group presentation**)

†**Sabot F**, Picault N, ElBaidouri M, Llauro C, Chaparro C, Piegu B, Roulin A, Guiderdoni E, Delabastide M, McCombie R, Panaud O : Transpositional landscape of rice genome revealed by Paired-End Mapping of high-throughput resequencing data. **The Plant Journal** 2011, 66 :241-246

Carrier G, Le Cunff L, Dereeper A, Legrand S, **Sabot F**, Bouchez O, Audeguin L, Boursiquot JM, This P : Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. **PLoS One** 2012, 7(3), e32973.

Jaligot E, Hooi WY, Debladis E, Richaud F, Beulé T, Collin M, Agbessi MDT, **Sabot F**, Garsmeur O, D'Hont A, Syed Alwee SSR, Rival A : DNA methylation and transcriptional activity of the *EgDEF1* gene and neighboring retrotransposons in *mantled* somaclonal variants of oil palm. **PLoS One** 2014, 9, e91896.

Nabholzl B, Sarah G, **Sabot F**, Ruiz M, Adam H, Nidelet S, Ghesquière A, Santoni S, David J, Glémin S : Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*O. glaberrima*). **Molecular Ecology** 2014, 23(9) :2210-2227

Orjuela J, **Sabot F**, Chéron S, Vigouroux Y, Adam H, Chrestin H, Sanni K, Lorieux M, Ghesquière A : An extensive analysis of the African rice genetic diversity through a global genotyping. **TAG Theoretical & Applied Genetics** 2014, 127(10) :2211-2223 - Co-premier auteur

Sabot F : *Tos17* rice element : incomplete but effective. **Mobile DNA** 2014, 5(1) :10

†Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, Kougbeadjo A, Maillol V, Martin G, **Sabot F**, Santoni S, Vigouroux Y, Couvreur T : Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. **Molecular Ecology Ressources** 2014

5.1.2 Soumis ou en préparation

3 papiers soumis, 3 en préparation. Sur ces 6, 1 en premier auteur, 4 en dernier auteur.

†Monat C, Tranchant-Dubreuil C, **Sabot F** : *LTRclassifier* : a pipeline and a website for fast and structural LTR retrotransposons classification in plants. Soumis à **BMC Bioinformatics**.

Ta KN, **Sabot F**, Adam H, Vigouroux Y, De Mita S, Ghesquière A, Do NV, Gantet P, Jouannic S : Time-shift of panicle meristem states in African rice species. Soumis à **Plant Physiol** - Co-premier Auteur.

Sarah G, Homa F, Pointet S, Contreras S, **Sabot F**, Nabholz B, Santoni S, Sauné L, Ardisson M, Chantret N, Sauvage C, Tregear J, Jourda C, Pot D, Vigouroux Y, ... and Manuel Ruiz : A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Soumis à **The Plant Journal**.

†Monat C, Tranchant-Dubreuil C, Ghesquière A,**Sabot F** : Two new assemblies of *O. glaberrima*. En preparation pour **BMC Genomics**.

†Monat C, Tranchant-Dubreuil C, Kougbeadjo A, Summo M, Farcy C, Agbessi M, **Sabot F** : PipelineNGS : A flexible ressource for NGS analysis. En préparation pour **BMC bioinformatics**.

Ndomassi T, Tranchant-Dubreuil C, Mouttet E, Banon S, **Sabot F** : Framboisine, a mini-cluster facility for Southern Countries. En préparation pour **Bioinformatics**.

5.2 Chapitres d'Ouvrage

2 chapitres d'ouvrages, un en premier auteur et un en dernier auteur

Sabot F, Schulman AH. Genomics of Transposable Elements in the *Triticeae*. **Genetics & Genomics of *Triticeae***, 2009, pp 387-405. Springer US ed

Chaparro C, **Sabot F**. Methods and software in NGS for TE analysis. **Methods Mol Biol**. 2012 ;859 :105-14.

5.3 Interventions en Conférence et Posters

5.3.1 Posters en tant que premier auteur

Conférences Nationales J'ai présenté 6 posters dans des conférences nationales, telles que les conférences CNET (Congrès National sur les Eléments Transposables) ou JOBIM (Journées Ouvertes en Biologie, Informatique et Modélisation)...

Conférences Internationales J'ai présenté 8 posters dans des conférences internationales, telles que ICTE (International Congress on Transposable Elements), Plant & Animal Genome, PlantGEM,...

5.3.2 Présentations en tant que premier auteur

Conférences Nationales J'ai effectué 10 conférences dans des congrès nationaux (CNET, IBC, Journées Omics de l'Université de Montpellier II,...), dont 3 invitées.

Conférences Internationales J'ai effectué 10 conférences dans des congrès internationaux (Plant & Animal Genome, ISRFG/International Symposium on Rice Functional Genomics, MetaRisks, Asilomar, Cotonou,...) dont 3 invitées.

5.3.3 Association à des posters et des présentations nationales et internationales

J'ai été associé à 10 posters internationaux et 10 conférences nationales ou internationales.

Bibliographie

[Argout et al., 2010] Argout, X., Salse, J., Aury, J.-M. M., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Guiltinan, M. J., Maximova, S. N., Others, Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Bérard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A. A. A. A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., Lanaud, C., and Berard, A. (2010). The genome of Theobroma cacao. *Nature genetics*, 43(2) :101–108.

[CAO et al., 2006] CAO, Q., LU, B.-R., XIA, H., RONG, J., SALA, F., SPADA, A., and GRASSI, F. (2006). Genetic Diversity and Origin of Weedy Rice (Oryza sativa f. spontanea) Populations Found in North-eastern China Revealed by Simple Sequence Repeat (SSR) Markers. *Ann Bot*, 98(6) :1241–1252.

[Carrier et al., 2012] Carrier, G., Le Cunff, L., Dereeper, A., Legrand, D., Sabot, F., Bouchez, O., Audeguin, L., Boursiquot, J.-M., and This, P. (2012). Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. *PloS one*, 7(3) :e32973.

[Chantret et al., 2004] Chantret, N., Cenci, A., Sabot, F., Anderson, O., and Dubcovsky, J. (2004). Sequencing of the Triticum monococcum hardness locus reveals good microcolinearity with rice. *Molecular Genetics and Genomics : MGG*, 271(4) :377–386.

[Chantret et al., 2008] Chantret, N., Salse, J., Sabot, F., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.-F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Leroy, P., Bernard, M., and Chalhoub, B. (2008). Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. *Journal of Molecular Evolution*, 66(2) :138–150.

[Chantret et al., 2005] Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.-F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P., and Chalhoub, B. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *The Plant Cell*, 17(4) :1033–1045.

[Chaparro and Sabot, 2012] Chaparro, C. and Sabot, F. (2012). Methods and Software in NGS for TE Analysis. In *Mobile Genetic Elements*, volume 859, pages 105–114. Springer.

[Chen et al., 2009] Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). BreakDancer : an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9) :677–681.

[Colas et al., 2008] Colas, I., Shaw, P., Prieto, P., Wanous, M., Spielmeyer, W., Mago, R., and Moore, G. (2008). Effective chromosome pairing requires chromatin remodeling at the onset of meiosis. *Proceedings of the National Academy of Sciences*, 105(16) :6075–6080.

[Dannull et al., 1994] Dannull, J., Surovoy, A., Jung, G., and Moelling, K. (1994). Specific binding of HIV-1 nucleocapsid protein to PSI RNA in vitro requires N-terminal zinc finger and flanking basic amino acid residues. *The EMBO Journal*, 13(7) :1525–1533.

[Davey et al., 2011] Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, 12(7) :499–510.

[Devos et al., 2002] Devos, K. M., Brown, J. K. M., and Bennetzen, J. L. (2002). Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Research*, 12(7) :1075–1079.

[Fei et al., 2013] Fei, Q., Xia, R., and Meyers, B. C. (2013). Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. *The Plant Cell Online*.

[Garavito et al., 2010] Garavito, A., Guyot, R., Lozano, J., Gavory, F., Samain, S., Panaud, O., Tohme, J., Ghesquiere, A., and Lorieux, M. (2010). A Genetic Model for the Female Sterility Barrier Between Asian and African Cultivated Rice Species. *Genetics*, 185(4) :1425–1440.

[Gridley et al., 2002] Gridley, H. E., Jones, M. P., and Wopereis-Pura, M. (2002). Development of new rice for Africa (NERICA) and participatory varietal selection. pages 12–15.

[Gross et al., 2010] Gross, B. L., Reagon, M., Hsu, S.-C., Caicedo, A. L., Jia, Y., Olsen, K. M., and Thurber, C. S. (2010). Molecular evolution of shattering loci in U.S. weedy rice. *Molecular Ecology*, 19(16) :3271–3284.

[Gutierrez et al., 2010] Gutierrez, A., Carabali, S., Giraldo, O., Martinez, C., Correa, F., Prado, G., Tohme, J., and Lorieux, M. (2010). Identification of a Rice stripe necrosis virus resistance locus and yield component QTLs using Oryza sativa x O. glaberrima introgression lines. *BMC Plant Biology*, 10(1) :6.

[Hajirasouliha et al., 2010] Hajirasouliha, I., Hormozdiari, F., Alkan, C., Kidd, J. M., Birol, I., Eichler, E. E., and Sahinalp, S. C. (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, 26(10) :btq152.

[Hirsch et al., 2014] Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell Online*, page tpc.113.119982.

[Hormozdiari et al., 2010] Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation VariationHunter : combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12) :i350–357.

[Jaligot et al., 2014] Jaligot, E., Hooi, W. Y., Debladis, E., Richaud, F., Beulé, T., Collin, M., Agbessi, M. D. T., Sabot, F., Garsmeur, O., D'Hont, A., Alwee, S. S. R. S., and Rival, A. (2014). DNA methylation and transcriptional activity of the EgDEF1 gene and neighboring retrotransposons in mantled somaclonal variants of oil palm. *PLoS ONE*, 9(3) :e91896.

[Johnson et al., 2009] Johnson, C., Kasprzewska, A., Tennessen, K., Fernandes, J., Nan, G.-L., Walbot, V., Sundaresan, V., Vance, V., and Bowman, L. H. (2009). Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome research*, 19(8) :1429–40.

[Kalendar et al., 2011] Kalendar, R., Flavell, A. J., Ellis, T. H. N., Sjakste, T., Moisy, C., and Schulman, A. H. (2011). Analysis of plant diversity with retrotransposon-based molecular markers. *Heredity*, 106(4) :520–530.

[Kalendar et al., 2000] Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A. H. (2000). Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences*, 97(12) :6603–6607.

[Kalendar et al., 2004] Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large Retrotransposon Derivatives : Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes. *Genetics*, 166(3) :1437–1450.

[Koren et al., 2012] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7) :693–700.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14) :1754 –1760.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079.

[Li et al., 2010] Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., Cann, H., Zhang, X., Li, S., Bolund, L., Kristiansen, K., Yang, H., Wang, J., and Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature biotechnology*, 28(1) :57–63.

[Lisitsyn, 1995] Lisitsyn, N. A. (1995). Representational difference analysis : finding the differences between genomes. *Trends in Genetics*, 11(8) :303–307.

[LONDO and SCHAAL, 2007] LONDO, J. P. and SCHAAL, B. A. (2007). Origins and population genetics of weedy red rice in the USA. *Molecular Ecology*, 16(21) :4523–4535.

[Mariac et al., 2014] Mariac, C., Scarcelli, N., Pouzadou, J., Barnaud, A., Billot, C., Faye, A., Kougbeadjo, A., Maillol, V., Martin, G., Sabot, F., Santoni, S., Vigouroux, Y., and Couvreur, T. L. P. (2014). Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular ecology resources*.

[Martin, 2011] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1) :pp. 10–12.

[MAYR, 1963] MAYR, E. (1963). Animal species and evolution. pages xiv + 797 pp.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M. a., and Others (2010). The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9) :1297–1303.

[Morgante et al., 2007] Morgante, M., De Paoli, E., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current opinion in plant biology*, 10(2) :149–55.

[Nabholz et al., 2014] Nabholz, B., Sarah, G., Sabot, F., Ruiz, M., Adam, H., Nidelet, S., Ghesquière, A., Santoni, S., David, J., and Glémin, S. (2014). Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (O. glaberrima). *Molecular Ecology*, pages n/a–n/a.

[Orjuela et al., 2014] Orjuela, J., Sabot, F., Chéron, S., Vigouroux, Y., Adam, H., Chrestin, H., Sanni, K., Lorieux, M., and Ghesquière, A. (2014). An extensive analysis of the African rice genetic diversity through a global genotyping. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*.

[Petit et al., 2009] Petit, J., Bourgeois, E., Stenger, W., Bès, M., Droc, G., Meynard, D., Courtois, B., Ghesquière, A., Sabot, F., Panaud, O., and Guiderdoni, E. (2009). Diversity of the Ty-1 copia retrotransposon Tos17 in rice (Oryza sativa L.) and the AA genome of the Oryza genus. *Molecular Genetics and Genomics : MGG*, 282(6) :633–652.

[Picault et al., 2009] Picault, N., Chaparro, C., Piegu, B., Stenger, W., Formey, D., Llauro, C., Descombin, J., Sabot, F., Lasserre, E., Meynard, D., Guiderdoni, E., and Panaud, O. (2009). Identification of an active LTR retrotransposon in rice. *The Plant Journal : For Cell and Molecular Biology*, 58(5) :754–765.

[Portères, 1962] Portères, R. (1962). Primary cradles of agriculture in the African continent. *Journal of African History*, 3 :195–210.

[Qi and Zhao, 2011] Qi, J. and Zhao, F. (2011). inGAP-sv : a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Research*, 39(Web Server issue) :W567–W575.

[Qiu et al., 2014] Qiu, J., Zhu, J., Fu, F., Ye, C.-Y., Wang, W., Mao, L., Lin, Z., Chen, L., Zhang, H., Guo, L., Qiang, S., Lu, Y., and Fan, L. (2014). Genome re-sequencing suggested a weedy rice origin from domesticated indica-japonica hybridization : a case study from southern China. *Planta*, pages 1–11.

[Roulin et al., 2009] Roulin, A., Piegu, B., Fortune, P. M., Sabot, F., D'Hont, A., Manicacci, D., and Panaud, O. (2009). Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. *BMC Evolutionary Biology*, 9 :58.

[Sabot, 2014] Sabot, F. (2014). Tos17 rice element : incomplete but effective. *Mobile DNA*, 5(1) :10.

[Sabot et al., 2005a] Sabot, F., Guyot, R., Wicker, T., Chantret, N., Laubin, B., Chalhoub, B., Leroy, P., Sourdille, P., and Bernard, M. (2005a). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Molecular Genetics and Genomics : MGG*, 274(2) :119–130.

[Sabot et al., 2006a] Sabot, F., Kalendar, R., Jääskeläinen, M., Wei, C., Tanskanen, J., and Schulman, A. H. (2006a). Retrotransposons : Metaparasites and Agents of Genome Evolution. *Israel Journal of Ecology and Evolution*, 52(3) :319–330.

[Sabot et al., 2011] Sabot, F., Picault, N., ElBAIDOURI, M., Llauro, C., Chaparro, C., Piegu, B., Roulin, A., Guiderdoni, E., Delabastide, M., McCOMBIE, R., and Panaud, O. (2011). Transpositional landscape of rice genome revealed by Paired-End Mapping of high-throughput resequencing data. *The Plant Journal*, pages no–no.

[Sabot and Schulman, 2006] Sabot, F. and Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants : a hitchhiker's guide to the genome. *Heredity*, 97(6) :381–388.

[Sabot and Schulman, 2007] Sabot, F. and Schulman, A. H. (2007). Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. *BMC Genomics*, 8 :247.

[Sabot and Schulman, 2009] Sabot, F. and Schulman, A. H. (2009). Genomics of transposable elements in the Triticeae. In *Genetics and Genomics of the Triticeae*, pages 387–405. Springer.

[Sabot et al., 2004] Sabot, F., Sourdille, P., and Bernard, M. (2004). Detecting specific repeated sequences in large, complex genomes by using representative difference analysis and double-probe verification. *Plant Molecular Biology Reporter*, 22(1) :91.

[Sabot et al., 2006b] Sabot, F., Sourdille, P., Chantret, N., and Bernard, M. (2006b). Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. *Genetica*, 128(1-3) :439–447.

[Sabot et al., 2005b] Sabot, F. F., Sourdille, P., and Bernard, M. (2005b). Advent of a new retrotransposon structure : the long form of the Veju elements. *Genetica*, 125(2-3) :325–332.

[Sakai et al., 2014] Sakai, H., Kanamori, H., Arai-Kichise, Y., Shibata-Hatta, M., Ebana, K., Oono, Y., Kurita, K., Fujisawa, H., Katagiri, S., Mukai, Y., Hamada, M., Itoh, T., Matsumoto, T., Katayose, Y., Wakasa, K., Yano, M., and Wu, J. (2014). Construction of Pseudomolecule Sequences of the aus Rice Cultivar Kasalath for Comparative Genomics of Asian Cultivated Rice. *DNA Research*.

[Sano, 1990] Sano, Y. (1990). The genic nature of gamete eliminator in rice. *Genetics*, 125(1) :183–191.

[Sarla and Mallikarjuna Swany, 2005] Sarla, N. and Mallikarjuna Swany, B. (2005). Oryza glaberrima : A source for the improvement of Oryza sativa. *Current Science*, 89(6) :955–963.

[Schatz et al., 2014] Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antonio, E., Ghiban, E., Wright, M. H., Chia, J.-m., Ware, D., McCouch, S. R., and McCombie, W. R. (2014). New whole genome de novo assemblies of three divergent strains of rice (O. sativa) documents novel gene space of aus and indica. *bioRxiv*.

[Schmitz and Ecker, 2012] Schmitz, R. J. and Ecker, J. R. (2012). Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends in plant science*.

[Semon et al., 2005] Semon, M., Nielsen, R., Jones, M. P., and McCouch, S. R. (2005). The population structure of African cultivated rice oryza glaberrima (Steud.) : evidence for elevated levels of linkage disequilibrium caused by admixture with O. sativa and ecological adaptation. *Genetics*, 169(3) :1639–1647.

[Slotkin and Martienssen, 2007] Slotkin, R. K. and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics*, 8(4) :272–85.

[Song et al., 2014] Song, B.-K., Chuah, T.-S., Tam, S. M., and Olsen, K. M. (2014). Malaysian weedy rice shows its true stripes : wild Oryza and elite rice cultivars shape agricultural weed evolution in Southeast Asia. *Molecular Ecology*, pages n/a–n/a.

[Tanskanen et al., 2007] Tanskanen, J. A., Sabot, F., Vicient, C., and Schulman, A. H. (2007). Life without GAG : The< i> BARE</i>-2 retrotransposon as a parasite's parasite. *Gene*, 390(1) :166–174.

[Tettelin et al., 2008] Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics : the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5) :472–477.

[Thurber et al., 2010] Thurber, C. S., Reagon, M., Gross, B. L., Olsen, K. M., Jia, Y., Caicedo, A. L., and Hsu, S.-C. (2010). Molecular evolution of shattering loci in U.S. weedy rice. *Molecular Ecology*, 19(16) :3271–3284.

[Vaughan et al., 2008] Vaughan, D. a., Lu, B.-R., and Tomooka, N. (2008). The evolving story of rice evolution. *Plant Science*, 174(4) :394–408.

[Wang et al., 2014] Wang, M., Yu, Y., Haberer, G., Marri, P. R., Fan, C., Goicoechea, J. L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J. S. S., Cossu, R. M., Maldonado, C., Chen, J., Lee, S., Sisneros, N., de Baynast, K., Golser, W., Wissotski, M., Kim, W., Sanchez, P., Ndjiondjop, M.-N., Sanni, K., Long, M., Carney, J., Panaud, O., Wicker, T., Machado, C. A., Chen, M., Mayer, K. F. X., Rounsley, S., and Wing, R. A. (2014). The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nature Genetics*, advance online publication.

[Wicker et al., 2008a] Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G. T. H., Graner, A., Ware, D., and Stein, N. (2008a). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC genomics*, 9(1) :518.

[Wicker et al., 2007] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12) :973–82.

[Wicker et al., 2008b] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2008b). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, 9(5) :414–414.

[Wicker et al., 2009] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2009). Reply : A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, 10(4) :276–276.

[Xia et al., 2009] Xia, Q., Guo, Y., Zhang, Z., Li, D. D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R., Cheng, T., Jiang, T., Becquet, C., Xu, X., Liu, C., Zha, X., Fan, W., Lin, Y., Shen, Y., Jiang, L., Jensen, J., Hellmann, I., Tang, S., Zhao, P., Xu, H., Yu, C., Zhang, G., Li, J. J., Cao, J., Liu, S., He, N., Zhou, Y., Liu, H., Zhao, J., Ye, C., Du, Z., Pan, G., Zhao, A., Shao, H., Zeng, W., Wu, P., Li, C., Pan, M., Yin, X., Wang, J. J. J., Zheng, H., Wang, W., Zhang, X., Li, S., Yang, H., Lu, C., Nielsen, R., Zhou, Z., and Xiang, Z. (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science (New York, N.Y.)*, 326(5951) :433–6.

[Zhao et al., 2011] Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., and McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, 2 :467.

Publications Principales